

Exact Lower Bounds for the Agnostic Probably-Approximately-Correct (PAC) Machine Learning Model

Aryeh Kontorovich and Iosif Pinelis

*Department of Computer Science
Ben-Gurion University
Beer Sheva, Israel 84105
e-mail: karyeh@cs.bgu.ac.il*

*Department of Mathematical Sciences
Michigan Technological University
Houghton, Michigan 49931-1295 U.S.A.
e-mail: ipinelis@mtu.edu*

Abstract: We provide an exact asymptotic lower bound on the minimax expected excess risk (EER) in the agnostic probably-approximately-correct (PAC) machine learning classification model. This bound is of the simple form $c_\infty/\sqrt{\nu}$ as $\nu \rightarrow \infty$, where $c_\infty = 0.16997\dots$ is a universal constant, $\nu = m/d$, m is the size of the training sample, and d is the Vapnik–Chervonenkis dimension of the hypothesis class. In the case when randomization of learning algorithms is allowed, we also provide an exact non-asymptotic lower bound on the minimax EER and identify minimax learning algorithms as certain maximally symmetric and minimally randomized “voting” procedures. It is shown that the differences between these asymptotic and non-asymptotic bounds, as well as the differences between these two bounds and the maximum EER of any learning algorithms that minimize the empirical risk, are asymptotically negligible, and all these differences are due to ties in the mentioned “voting” procedures. A few easy to compute non-asymptotic lower bounds on the minimax EER are also obtained, which are shown to be close to the exact asymptotic lower bound $c_\infty/\sqrt{\nu}$ even for rather small values of the ratio $\nu = m/d$. As an application of these results, we substantially improve existing lower bounds on the tail probability of the excess risk. Among the tools used are Bayes estimation and apparently new identities and inequalities for binomial distributions.

AMS 2010 subject classifications: Primary 68T05, 62C20, 62C10, 62C12, 62G20, 62H30; secondary 62G10, 62C20, 91A35, 60C05.

Keywords and phrases: PAC learning theory, classification, generalization error, minimax decision rules, Bayes decision rules, empirical estimators, binomial distribution.

Contents

1	Introduction	2
2	Results: statements and discussion	6
3	Proofs	16

* June 30, 2016

A	Identities and inequalities for binomial distributions: details concerning the function bayes	25
B	Details concerning $c_{m,d}^{\text{LB}}$	28
C	Index of symbols	28
	References	28

1. Introduction

The Probably Approximately Correct (PAC) model aims at providing a clean, plausible and minimalistic abstraction of the supervised learning process [20, 19]. The theory has been refined into a number of variants, and in this paper we are concerned with the one most commonly appearing in modern literature, *agnostic PAC* [5, 8, 7].

Let \mathcal{X} be an arbitrary nonempty set. An element x of \mathcal{X} can be thought of as a possibly incomplete description of a corresponding real object, which latter is a member of a certain population of objects; say, x may be the pair (height, weight) of a person. The objective is to classify the elements of the set \mathcal{X} (referred to as the *instance space*) into two classes, by attaching a label 1 or -1 to each $x \in \mathcal{X}$; of course, any such binary classification of the descriptions x will induce a binary classification of the underlying objects. Let $\mathcal{Y} := \{-1, 1\}$, the set of labels. Then a possible classification may be identified with a map $h: \mathcal{X} \rightarrow \mathcal{Y}$, called a *hypothesis*. Usually, hypotheses are restricted to be elements of a specified subset \mathcal{H} of the set $\mathcal{Y}^{\mathcal{X}}$ of all maps of \mathcal{X} to \mathcal{Y} ; this subset \mathcal{H} is called the *hypothesis class*. For example, if \mathcal{X} is a subset of a Euclidean space \mathcal{E} with an inner product $\mathcal{E}^2 \ni (x, y) \mapsto x \cdot y$, then hypothesis class \mathcal{H} may be the set of all so-called perceptrons $h_{w,\theta}$, given by the formula $h_{w,\theta}(x) = \text{sgn}(w \cdot x - \theta)$ for some $w \in \mathcal{E}$ and $\theta \in \mathbb{R}$ and all $x \in \mathcal{X}$, where

$$\text{sgn } u := 2\mathbf{I}\{u \geq 0\} - 1$$

for real u and $\mathbf{I}\{\cdot\}$ is the indicator function.

Since a description x may be incomplete, it may be not enough to identify the corresponding object, and so, different descriptions may have different frequencies in the population of objects. Incompleteness of descriptions may also make it impossible to classify an object with accuracy and certainty based only on the corresponding description; indeed, two different objects with the same incomplete description x may actually belong to different classes. Therefore, it is natural to assume that there exists a true (but unknown to us) probability distribution, say D , on the set $\mathcal{X} \times \mathcal{Y}$ of all pairs (x, y) with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. To avoid tedious matters of measurability, let us just assume that the set \mathcal{X} is finite.

In the agnostic PAC model, considered in this paper, it is assumed that the distribution D may be of completely arbitrary form, and the only information about it is provided to us by the “sample” values of a *labeled sample* $(X_1^D, Y_1^D), \dots, (X_m^D, Y_m^D)$ of m independent copies of a random pair

(X^D, Y^D) ; here and in what follows, the superscript indicates the distribution of the random pair.

In the so-called restricted (also: realizable or consistent) model (see e.g. [2]), in contrast with the agnostic one, it is assumed that the support of the unknown distribution D is the graph of a function from \mathcal{X} to \mathcal{Y} . This corresponds to the assumption, in informal terms, that the descriptions x of the objects are sufficiently complete — so that, to adequately classify an object, it is enough to classify its description. For example, in a restricted model, persons may be classified as obese based just on their (height, weight) description.

However, let us turn back to the agnostic PAC model. In such a model, even the complete knowledge of the true distribution D would not make an error-free classification possible, again because the descriptions x may be incomplete. Indeed, the error probability for a hypothesis $h \in \mathcal{H}$ is

$$\text{err}(h, D) := \mathbb{P}(h(X^D) \neq Y^D). \quad (1.1)$$

It should now be clear that in general the least possible classification error

$$\text{err}_{\min}(D) := \min_{h \in \mathcal{H}} \text{err}(h, D) \quad (1.2)$$

will indeed usually be strictly greater than 0, even when the true distribution D is known.

Recall that in the agnostic PAC model, considered here, the only information about the unknown distribution D is provided by the values of the sequence of m independent random pairs

$$Z_m^D := ((X_1^D, Y_1^D), \dots, (X_m^D, Y_m^D)). \quad (1.3)$$

Therefore, the available “learning” strategies are the mappings

$$L: (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathcal{H},$$

called *learning algorithms*.

Let h_D denote any minimizer of $\text{err}(h, D)$ over $h \in \mathcal{H}$. Of course, h_D is unknown, since the distribution D is unknown. However, it may be reasonable to use the plug-in estimator $h_{\hat{D}_m}$ of h_D , obtained by substituting for D the empirical distribution $\hat{D}_m = \hat{D}_m(z_m)$ based on a “realization”

$$z_m := ((x_1, y_1), \dots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m \quad (1.4)$$

of the “random sample” Z_m from the distribution D . That is,

$$h_{\hat{D}_m} = L_{\text{emp}}(Z_m^D),$$

where L_{emp} is any learning algorithm such that for each given sequence $z_m \in (\mathcal{X} \times \mathcal{Y})^m$, the corresponding value $L_{\text{emp}}(z_m)$ of L_{emp} is a minimizer in $h \in \mathcal{H}$ of the “empirical risk”

$$\text{err}(h, \hat{D}_m) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{h(x_i) \neq y_i\}.$$

Such a minimizer need not be unique, and so, the “empirical minimization” learning algorithm L_{emp} does not have to be unique.

Putting aside the algorithmic question of *how* to efficiently select a minimizer $h_{\hat{D}_m}$ of the empirical risk, the statistical question of *sample complexity* remains. At first glance, what one might wish to know is how the performance of $h_{\hat{D}_m}$ improves with increasing sample size m . Unfortunately, the answer might be trivial: for example, when $\mathbf{P}(Y^D = 1 | X^D) = \mathbf{P}(Y^D = -1 | X^D) = \frac{1}{2}$, we have $\text{err}(h, D) = \frac{1}{2}$ for all $h \in \mathcal{H}$. Instead, a nontrivial question to ask is how well the empirical risk minimizer $h_{\hat{D}_m}$ performs compared to the best possible hypothesis, h_D — that is, how large the *excess risk* $\Delta(h_{\hat{D}_m}, D)$ is, where

$$\Delta(h, D) := \text{err}(h, D) - \text{err}_{\min}(D) = \text{err}(h, D) - \text{err}(h_D, D). \quad (1.5)$$

This question has been to a large extent resolved. In particular, Theorem 4.9 from [2] (slightly restated here) provides the following upper bound on the tail probabilities for the excess risk.

Theorem A. *There is a universal real constant $c > 0$ such that for all finite sets \mathcal{X} , all distributions D on $\mathcal{X} \times \mathcal{Y}$, all sample sizes m , and all hypothesis classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ of VC dimension d , we have*

$$\mathbf{P}(\Delta(L_{\text{emp}}(Z_m^D), D) \geq cu) \leq \exp\{-(mu^2 - d)_+\} \quad (1.6)$$

for all real $u \geq 0$, where $z_+ := 0 \vee z$ for real z .

See [2] for an account of the intermediate steps leading up to the seminal and highly non-trivial result presented in Theorem A; notable milestones here include the papers [20, 17, 6, 9].

Recall that the VC dimension (that is, the Vapnik–Chervonenkis dimension) of a set $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is the largest nonnegative integer k such that there is a subset of \mathcal{X} of cardinality k that is shattered by \mathcal{H} ; and a subset \mathcal{X}_0 of \mathcal{X} is said to be shattered by \mathcal{H} if the set of the restrictions to \mathcal{X}_0 of all the functions $h \in \mathcal{H}$ coincides with the entire set $\mathcal{Y}^{\mathcal{X}_0}$ of all functions from \mathcal{X}_0 to \mathcal{Y} .

In what follows d will always denote the VC dimension of \mathcal{H} . The case $d = 0$ may occur only if the cardinality of \mathcal{H} is at most 1, so that there is at most one hypothesis to choose. This trivial case will be excluded in the sequel; that is, we shall assume that

$$d \geq 1.$$

Then, in particular, one can introduce the fundamental ratio

$$\nu := \nu_{m,d} := m/d \quad (1.7)$$

of the sample size m to the VC dimension d .

Lower bounds nearly matching, up to constant factors, the upper bound given in Theorem A are also known. The one with the apparently best currently known numerical constants was given in [2, Theorem 5.2], which can be restated as follows.

Theorem B. *If $\nu = m/d \geq 64^2/320 = 12.8$, then for every set \mathcal{X} and a hypothesis classe $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ of VC dimension d , and any learning algorithm $L: (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathcal{H}$, there is a distribution D on $\mathcal{X} \times \mathcal{Y}$ such that*

$$\mathbb{P} \left(\Delta(L(Z_m^D), D) > \frac{1}{\sqrt{320\nu}} \right) \geq \frac{1}{64}. \quad (1.8)$$

This lower bound is also the culmination of a notable historical development [20, 4, 15], detailed in [2]. It is stated there: “The bounds in [Theorem 5.2 of [2]] improve (by constants) all previous bounds” — at the time, and apparently also to-date.

Remark 1.1. In [14, Section 28.2.2] a much better constant factor, $1/8$, was claimed in place of $1/320$. However, there is a mistake in the calculation; the actual value of the constant furnished by the proof is 512 .

Introduce the *expected excess risk*

$$\mathfrak{R}(L, D) := \mathfrak{R}_m(L, D) := \mathbb{E} \Delta(L(Z_m^D), D). \quad (1.9)$$

Let $\mathfrak{H}_{\mathcal{X},d}$ denote the set of all hypothesis classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ of VC dimension d . Let then

$$\begin{aligned} c_{m,d}^{\text{UB}} &:= \sqrt{\nu_{m,d}} \sup_{\mathcal{X}} \sup_{\mathcal{H} \in \mathfrak{H}_{\mathcal{X},d}} \inf_L \sup_D \mathfrak{R}_m(L, D), \\ c_{m,d}^{\text{LB}} &:= \sqrt{\nu_{m,d}} \inf_{\mathcal{X}} \inf_{\mathcal{H} \in \mathfrak{H}_{\mathcal{X},d}} \inf_L \sup_D \mathfrak{R}_m(L, D), \end{aligned} \quad (1.10)$$

where $\sup_{\mathcal{X}}$ and $\inf_{\mathcal{X}}$ are taken over all finite sets \mathcal{X} ; \inf_L is taken over all learning algorithms $L: (\mathcal{X} \times \mathcal{Y})^k \rightarrow \mathcal{H}$; and \sup_D is taken over all distributions D on $\mathcal{X} \times \mathcal{Y}$. The quantity $\inf_L \sup_D \mathfrak{R}_m(L, D)$ may be referred to as the *minimax expected excess risk*.

Integrating both sides of inequality (1.6) in $u \geq 0$, one sees that

$$c_{m,d}^{\text{UB}} := \sup_{m,d} c_{m,d}^{\text{UB}} \leq \sup_{m,d} \sqrt{\nu_{m,d}} \sup_{\mathcal{X}} \sup_{\mathcal{H} \in \mathfrak{H}_{\mathcal{X},d}} \sup_D \mathfrak{R}_m(L_{\text{emp}}, D) < \infty, \quad (1.11)$$

where $\sup_{m,d}$ is taken over all natural m and d ; an exact calculation of c^{UB} seems to be beyond the reach of current methods, which only yield loose estimates.

It is also clear that inequality (1.8) implies

$$c_{\nu \geq 12.8}^{\text{LB}} > \frac{1}{64\sqrt{320}} = 0.000873\dots > 0, \quad (1.12)$$

where

$$c_{\nu \geq \nu_*}^{\text{LB}} := \inf \{ c_{m,d}^{\text{LB}} : \nu_{m,d} \geq \nu_* \}$$

for any real $\nu_* > 0$. A remarkable fact that follows from (1.11) and (1.12) is that

$$0 < \liminf_{m/d \rightarrow \infty} c_{m,d}^{\text{LB}} \leq \limsup_{m/d \rightarrow \infty} c_{m,d}^{\text{UB}} < \infty;$$

that is, the upper and lower bounds on the minimax expected excess risk are of the same order of magnitude.

Establishing an appropriate lower bound on the expected excess risk $\mathfrak{R}(L, D)$ was the crucial part of the proof of Theorem B. Actually, following the lines of the proof of [2, Theorem 5.2], one can see that the lower bound $\frac{1}{64\sqrt{320}} = 0.000873\dots$ in (1.12) can be improved to 0.0675 for large enough $\nu = m/d$ — cf. Appendix B.

In this paper, we present optimal lower bounds on the minimax expected excess risk, which cannot be further improved. In particular, we shall show (in Theorem 2.4) that

$$c_{m,d}^{\text{LB}} \rightarrow c_\infty := \max_{z>0} \frac{z}{2} (1 - \text{erf}(z/\sqrt{2})) = 0.16997\dots \quad (1.13)$$

whenever m and d vary in such a way that $\nu_{m,d} \rightarrow \infty$; here, as usual, erf denotes the Gauss error function, given by the formula $\text{erf}(u) := \frac{2}{\sqrt{\pi}} \int_0^u e^{-t^2} dt$ for real $u \geq 0$. Moreover, it will be noted (in Remark 2.7) that the values of $c_{m,d}^{\text{LB}}$ are rather close to the limit c_∞ already for rather small values of $\nu_{m,d}$. Furthermore, in Theorem 2.1 we shall provide an exact expression for the minimax expected excess risk when randomization of learning algorithms is allowed; it will also be shown there that the effect of this randomization is asymptotically negligible and is entirely explained by ties in a certain “voting” procedure.

As mentioned above, the proof of [2, Theorem 5.2] was based on a lower bound on the expected excess risk $\mathfrak{R}(L, D)$. Accordingly, using our improved (and optimal) lower bounds on $\mathfrak{R}(L, D)$, one can substantially improve the constants in [2, Theorem 5.2] or, equivalently, in Theorem B of the present paper, in particular as follows.

Theorem 1.2.

- (i) Keeping the constants 12.8 and 320 in Theorem B in place, one can improve the lower bound $\frac{1}{64} \approx 0.0156$ on the tail probability in (1.8) to 0.238.
- (ii) Keeping the constants 12.8 and $\frac{1}{64}$ in Theorem B in place, one can improve the constant 320 in (1.8) to 41.3.
- (iii) If the restriction $\nu \geq 12.8$ in Theorem B is relaxed to $\nu \geq 3$, then the improved values 0.238 and 41.3 of the constants get only slightly worse: 0.227 and 49.6, respectively.

A few words on the organization of the rest of this paper: The main results are stated and discussed in Section 2. All necessary proofs are given in Section 3, with more technical parts deferred further, to Appendices A and B. An index of symbols used in this paper is given in Appendix C.

2. Results: statements and discussion

First here, let us introduce some notation and conventions to be used in what follows.

Let $0^0 := 1$.

For any α and β in $\mathbb{Z} \cup \{\infty\}$, let $\overline{\alpha, \beta} := \{i \in \mathbb{Z} : \alpha \leq i \leq \beta\}$. For any $m \in [0, \infty]$, let $[m] := \overline{1, m}$. In particular, $[0] = \emptyset$.

As usual, for any two sets S and T , let S^T denote the set of all maps from T to S .

For any set A and any $k \in \overline{0, \infty}$ we identify the k -tuples $v = (v_1, \dots, v_k) \in A^k$ with functions $v: [k] \rightarrow A$, by the formula $v(x) := v_x$ for all $x \in [k]$; thus, we identify the set A^k of k -tuples with the set $A^{[k]}$ of functions. So, we use notations $v(x)$ and v_x interchangeably. We shall also identify a function with its graph.

As usual, the sum of the empty family of elements of a linear space is defined as the zero element of that space.

The new results obtained in this paper all concern the lower bound $c_{m,d}^{\text{LB}}$, defined in (1.10), on the minimax expected excess risk times the factor $\sqrt{\nu_{m,d}}$, including the limit behavior of $c_{m,d}^{\text{LB}}$ as $\nu_{m,d} = m/d \rightarrow \infty$.

Take any finite set \mathcal{X} . Take then any $\mathcal{H} \in \mathfrak{H}_{\mathcal{X},d}$; that is, $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is a hypothesis class of VC dimension d . Let now $\tilde{\mathcal{X}}$ be any subset of \mathcal{X} of cardinality d such that $\tilde{\mathcal{X}}$ is shattered by \mathcal{H} . Clearly, $\sup_{\tilde{D}} \mathfrak{R}(L, D) \geq \sup_{\tilde{D}} \mathfrak{R}(L, \tilde{D})$, where $\sup_{\tilde{D}}$ is taken over all distributions \tilde{D} on $\mathcal{X} \times \mathcal{Y}$ with support contained in the set $\tilde{\mathcal{X}} \times \mathcal{Y}$.

Thus, the defining expression for $c_{m,d}^{\text{LB}}$ in (1.10) can be simplified:

$$c_{m,d}^{\text{LB}} = \sqrt{\nu_{m,d}} \inf_L \sup_D \mathfrak{R}_m(L, D), \quad (2.1)$$

where from now on it will be assumed that

$$\mathcal{X} = [d] \quad \text{and} \quad \mathcal{H} = \mathcal{Y}^{\mathcal{X}} = \{-1, 1\}^{[d]}$$

and, for these particular \mathcal{X} and \mathcal{H} , \inf_L is still taken over all learning algorithms $L: (\mathcal{X} \times \mathcal{Y})^k \rightarrow \mathcal{H}$ and \sup_D is still taken over all distributions D on $\mathcal{X} \times \mathcal{Y}$. Accordingly, from now on we shall use \mathcal{X} and \mathcal{H} interchangeably with $[d]$ and $\mathcal{Y}^{\mathcal{X}} = \{-1, 1\}^{[d]}$, respectively.

Note next that any distribution D on $\mathcal{X} \times \mathcal{Y}$ is completely characterized by the two maps, say $p = p_D: \mathcal{X} \rightarrow [0, 1]$ and $\gamma = \gamma_D: \mathcal{X} \rightarrow [-1, 1]$, such that

$$\mathbb{P}(X^D = x, Y^D = y) = D(\{(x, y)\}) = p(x) \left(\frac{1}{2} + \frac{y\gamma(x)}{2} \right) \quad (2.2)$$

for $x \in \mathcal{X} = [d]$ and $y \in \mathcal{Y} = \{-1, 1\}$. Clearly then, one must have $p_D(x) = \mathbb{P}(X^D = x)$ for all $x \in \mathcal{X}$ and $\gamma_D(x) = 2\mathbb{P}(Y = 1|X = x) - 1$ for all $x \in \mathcal{X}$ with $p_D(x) \neq 0$; if $p_D(x) = 0$ for some $x \in \mathcal{X}$, then the value of $\gamma_D(x)$ can be chosen arbitrarily in $[-1, 1]$. So, the distribution of the r.v. X^D is completely characterized by the map $p = p_D$. Therefore, in what follows let us write $D = D_{p,\gamma}$ if $p_D = p$ and $\gamma_D = \gamma$, and, in the case when $D = D_{p,\gamma}$, let us simply write $X^p, Y^{p,\gamma}, X_i^p, Y_i^{p,\gamma}$ instead of X^D, Y^D, X_i^D, Y_i^D (respectively), assuming that the random pairs $(X_1^p, Y_1^{p,\gamma}), \dots, (X_m^p, Y_m^{p,\gamma})$ are independent copies of the random pair $(X^p, Y^{p,\gamma}) = (X^D, Y^D)$; let us then also write

$$Z_m^{p,\gamma} := Z_m^D := ((X_1^p, Y_1^{p,\gamma}), \dots, (X_m^p, Y_m^{p,\gamma})) \quad (2.3)$$

(cf. (1.3)).

Take next any $h \in \mathcal{H} = \{-1, 1\}^{[d]}$. Then, recalling (1.1) and (2.2) and noting that $\mathbf{I}\{h(X) = -1\} = 1 - \mathbf{I}\{h(X) = 1\}$, one has

$$\begin{aligned} \text{err}(h, D_{p,\gamma}) &= \mathbf{P}(h(X^p) \neq Y^{p,\gamma}) \\ &= \mathbf{P}(h(X^p) = 1, Y^{p,\gamma} = -1) + \mathbf{P}(h(X^p) = -1, Y^{p,\gamma} = 1) \\ &= \sum_{x=1}^d p_x \left(\frac{1}{2} - \frac{\gamma_x}{2} \right) \mathbf{I}\{h(x) = 1\} + \sum_{x=1}^d p_x \left(\frac{1}{2} + \frac{\gamma_x}{2} \right) \mathbf{I}\{h(x) = -1\} \\ &= \frac{1}{2} \mathbf{E}(1 - \gamma(X^p)) \mathbf{I}\{h(X^p) = 1\} + \frac{1}{2} \mathbf{E}(1 + \gamma(X^p)) \mathbf{I}\{h(X^p) = -1\} \\ &= \frac{1}{2} \mathbf{E}(1 + \gamma(X^p)) - \mathbf{E} \gamma(X^p) \mathbf{I}\{h(X^p) = 1\}. \end{aligned}$$

It is now clear that a minimizer of $\text{err}(h, D_{p,\gamma})$ over all $h \in \{-1, 1\}^{[d]}$ is the function $h_\gamma \in \{-1, 1\}^{[d]}$ given by the formula

$$h_\gamma(x) := \text{sgn } \gamma_x \quad (2.4)$$

for $x \in [d]$. More generally, a function $h \in \{-1, 1\}^{[d]}$ is a minimizer of $\text{err}(h, D_{p,\gamma})$ for given p and γ if and only if $h(x) = \text{sgn } \gamma_x [= \text{sgn}(p_x \gamma_x)]$ for all $x \in \mathcal{X}$ such that $p_x \gamma_x \neq 0$ (if $p_x \gamma_x = 0$ for some $x \in \mathcal{X}$, then the value $h(x)$ of a minimizer h at this point can be chosen arbitrarily in the set $\{-1, 1\}$).

Moreover, the excess risk (relative to $D_{p,\gamma}$) of h over h_γ is

$$\begin{aligned} \Delta(h, D_{p,\gamma}) &= \text{err}(h, D_{p,\gamma}) - \text{err}(h_\gamma, D_{p,\gamma}) \\ &= \mathbf{E} \gamma(X^p) (\mathbf{I}\{h_\gamma(X^p) = 1\} - \mathbf{I}\{h(X^p) = 1\}) \\ &= \mathbf{E} |\gamma(X^p)| \mathbf{I}\{h(X^p) \neq h_\gamma(X^p)\} \\ &= \sum_{x=1}^d p_x |\gamma_x| \mathbf{I}\{h(x) \neq h_\gamma(x)\}. \end{aligned} \quad (2.5)$$

Replacing now the unknown true distribution $D = D_{p,\gamma}$ by the empirical distribution $\hat{D}_m = \hat{D}_m((x_1, y_1), \dots, (x_m, y_m))$ for $((x_1, y_1), \dots, (x_m, y_m)) =: z \in (\mathcal{X} \times \mathcal{Y})^m$, one sees that a function $h \in \{-1, 1\}^{[d]}$ is a minimizer of $\text{err}(h, \hat{D}_m)$ for the given “sample” z if and only if $h(x) = \text{sgn } \widehat{p}\gamma_x$ for all $x \in \mathcal{X}$ such that $\widehat{p}\gamma_x \neq 0$, where

$$\widehat{p}\gamma_x := \frac{1}{m} \sum_{i=1}^m y_i \mathbf{I}\{x_i = x\};$$

if $\widehat{p}\gamma_x = 0$ for some $x \in \mathcal{X}$, then the value $h(x)$ of a minimizer h (of $\text{err}(h, \hat{D}_m)$) at this point can be chosen arbitrarily in the set $\{-1, 1\}$. Thus, all the learning algorithms L_{ERM} that are minimizers of the empirical risk are given by the formula

$$L_{\text{ERM}}(z_m)(x) := L_{m,d;\text{ERM}}(x) \begin{cases} := \text{sgn } v_x & \text{if } v_x \neq 0, \\ \in \{-1, 1\} & \text{if } v_x = 0 \end{cases} \quad (2.6)$$

for all $z_m \in (\mathcal{X} \times \mathcal{Y})^m$ and $x \in \mathcal{X} = [d]$, where

$$v_x := v_x(z_m) := \sum_{i=1}^m y_i \mathbf{I}\{x_i = x\} = m \widehat{p\gamma}_x. \quad (2.7)$$

Formula (2.6) states that the empirical risk is minimized when the value $y \in \{-1, 1\}$ assigned by the learning algorithm at point x based on the “sample” z_m is decided by the majority vote $v_x = v_x(z_m)$ “at x ”, with the “voting” restricted to the pairs (x_i, y_i) with $x_i = x$; if there is a tie (no majority) at x , then a value $y \in \{-1, 1\}$ at x is chosen arbitrarily.

To decrease the risk and also be able to fully use the power of decision theory, one may randomize learning algorithms, which may also better reflect the spirit of the agnostic model (in contrast with the restricted one). Of course, randomized decision rules, in particular randomized tests, are quite common and useful in statistics. A convenient way to define a randomized learning algorithm L is by allowing its value (which is a function in \mathcal{H}) to depend, not only on the nonrandom “sample” $z_m = ((x_1, y_1), \dots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ as in (1.4), but also on the value u of another r.v., say U , which is (say) uniformly distributed on the interval $[-1, 1]$ and independent of the random “sample” $Z_m^D = ((X_1^D, Y_1^D), \dots, (X_m^D, Y_m^D))$ as in (1.3). Thus, a randomized learning algorithm L will be understood as a Borel-measurable map from $(\mathcal{X} \times \mathcal{Y})^m \times [-1, 1]$ to \mathcal{H} .

Let $\mathcal{L}_{\text{rand}} = \mathcal{L}_{\text{rand}, m}$ and $\mathcal{L} = \mathcal{L}_m$ denote, respectively, the set of all randomized learning algorithms and the set of all non-randomized ones. The definition (1.9) of the expected excess risk (for $L \in \mathcal{L}$) is naturally extended as follows:

$$\mathfrak{R}(L, D) := \mathfrak{R}_m(L, D) := \mathbb{E} \Delta(L(Z_m^D, U), D) \quad (2.8)$$

for $L \in \mathcal{L}_{\text{rand}}$; it then follows by (2.5) that

$$\mathfrak{R}(L; p, \gamma) := \mathfrak{R}(L, D_{p, \gamma}) = \sum_{x=1}^d p_x |\gamma_x| \mathbb{P} \left(L(Z_m^{p, \gamma}, U)(x) \neq h_\gamma(x) \right). \quad (2.9)$$

Of particular importance will be the following “maximally symmetric” and “minimally randomized” version of the learning algorithms L_{ERM} that are minimizers of the empirical risk (cf. (2.6)):

$$L_{\text{ERM}}^*(z_m, u)(x) := L_{m, \text{ERM}}^*(z_m, u)(x) := \begin{cases} \text{sgn } v_x & \text{if } v_x \neq 0, \\ y_{i(x)} & \text{if } v_x = 0 \text{ but } n_x \neq 0, \\ \text{sgn } u & \text{if } n_x = 0 \end{cases} \quad (2.10)$$

for $(z_m, u) \in (\mathcal{X} \times \mathcal{Y})^m \times [-1, 1]$, where

$$n_x := n_x(z_m) := \sum_{i=1}^m \mathbf{I}\{x_i = x\} \quad \text{and} \quad i_x := i_x(z_m) := \min\{i \in [m] : x_i = x\}. \quad (2.11)$$

That is, the choice of the value of $L_{\text{ERM}}^*(z_m, u)(x)$ in $\mathcal{Y} = \{-1, 1\}$ is decided by the majority vote “at x ” if there is a majority there; otherwise, the value $L_{\text{ERM}}^*(z_m, u)(x)$ is the same as that of the first voter that appeared “at x ” if any one did; finally, if no one arrived to vote “at x ”, then the value is decided by a flip of a fair coin, the flip being independent of any voters. Thus, randomization according to the learning algorithm L_{ERM}^* occurs only if no one shows up for voting at some location $x \in \mathcal{X}$. Yet, this minimal (and, one may argue, quite natural) randomization is enough to make L_{ERM}^* a winner (that is, a minimax learning algorithm) against all randomized (and non-randomized) learning algorithms. A precise formulation of this thesis is contained in

Theorem 2.1. *Take any $m \in \overline{0, \infty}$. Then*

$$\begin{aligned} \inf_{L \in \mathcal{L}_{\text{rand}}} \sup_D \mathfrak{R}_m(L, D) &= \sup_D \mathfrak{R}_m(L_{\text{ERM}}^*, D) \\ &= B(m, d) := \sup_{p, \gamma} \sum_{x=1}^d p_x |\gamma_x| \mathbf{E} \text{ bayes}(N_x^p, |\gamma_x|), \end{aligned} \quad (2.12)$$

where $\sup_{p, \gamma}$ is taken over all pairs of functions $p \in [0, 1]^{[d]}$ such that $\sum_{x=1}^d p_x = 1$ and $\gamma \in [-1, 1]^{[d]}$, N_x^p is a r.v. with the binomial distribution with parameters m and p_x ,

$$\text{bayes}(k, b) := \frac{1}{2} (1 - s_k(b)), \quad (2.13)$$

$$s_k(b) := |\mathbf{P}(V_k^b > 0) - \mathbf{P}(V_k^{-b} > 0)|, \quad (2.14)$$

and V_k^b is a r.v. with the binomial distribution with parameters k and $\frac{1}{2}(1+b)$, for $k \in \overline{0, \infty}$ and $b \in [-1, 1]$; in particular, in accordance with the convention about the zero sum of the empty family, $V_0^b = 0$ and hence $s_0(b) = 0$ and $\text{bayes}(0, b) = \frac{1}{2}$. Moreover, for each pair of functions p and γ as described above,

$$\mathfrak{R}_m(L_{\text{ERM}}^*, D_{p, \gamma}) = \sum_{x=1}^d p_x |\gamma_x| \mathbf{E} \text{ bayes}(N_x^p, |\gamma_x|), \quad (2.15)$$

which does not depend on $\text{sgn } \gamma := (\text{sgn } \gamma_1, \dots, \text{sgn } \gamma_d)$.

Remark 2.2. It was observed in [3] that the value of $s_k(b)$ may also be expressed as the variational distance $\|\mu_k^b - \mu_k^{-b}\|_{\text{tv}}$, where $\mu_k^{\pm b}$ stands for the distribution of $V_k^{\pm b}$.

The use of the symbol bayes in (2.12) is a reflection of the fact that the minimax learning algorithm L_{ERM}^* is a Bayes one with respect to a certain prior distribution on the set of all distributions D on $\mathcal{X} \times \mathcal{Y}$; see the beginning of the proof of Theorem 2.1 in Section 3 for details on this. Formula (2.15) means that the learning algorithm L_{ERM}^* has an important risk-equalizing property — which actually makes the Bayes decision rule L_{ERM}^* minimax.

Remark 2.3. It is clear from (2.13)–(2.14) that $\text{bayes} \leq \frac{1}{2}$. Hence, by (2.12), $\inf_{L \in \mathcal{L}_{\text{rand}}} \sup_D \mathfrak{R}_m(L, D) \leq \frac{1}{2}$.

It turns out, as may be expected, that the effect of the randomization of learning algorithms is asymptotically negligible whenever $\nu = m/d \rightarrow \infty$; that is, the difference $\inf_{L \in \mathcal{L}} \sup_D - \inf_{L \in \mathcal{L}_{\text{rand}}} \sup_D$ is asymptotically negligible compared with the “non-randomized” minimax expected excess risk $\inf_{L \in \mathcal{L}} \sup_D$. Moreover, all the learning algorithms of the form L_{ERM} as in (2.6) that are minimizers of the empirical risk are asymptotically minimax. These facts — along with the asymptotics of the minimax risk — are presented in

Theorem 2.4. *For each pair (m, d) of natural numbers, choose any learning algorithm of the form $L_{m,d;\text{ERM}}$, as in (2.6). Then*

$$\begin{aligned} \frac{c_\infty}{\sqrt{\nu_{m,d}}} &\sim \inf_{L \in \mathcal{L}_{\text{rand}}} \sup_D \mathfrak{R}_m(L, D) \leq \inf_{L \in \mathcal{L}} \sup_D \mathfrak{R}_m(L, D) \\ &\leq \sup_D \mathfrak{R}_m(L_{m,d;\text{ERM}}, D) \sim \frac{c_\infty}{\sqrt{\nu_{m,d}}} \end{aligned} \quad (2.16)$$

whenever $\nu_{m,d} = m/d \rightarrow \infty$, where $c_\infty = 0.16997\dots$ as in (1.13). Moreover,

$$\begin{aligned} 0 &\leq \sup_D \mathfrak{R}_m(L_{m,d;\text{ERM}}, D) - \inf_{L \in \mathcal{L}_{\text{rand}}} \sup_D \mathfrak{R}_m(L, D) \\ &\leq \frac{1}{2} \sup_{p, \gamma} \sum_{x=1}^d p_x |\gamma_x| \mathbf{P}(V_x^{p, \gamma} = 0) = O\left(\frac{1}{\nu_{m,d}}\right) = o\left(\frac{1}{\sqrt{\nu_{m,d}}}\right), \end{aligned} \quad (2.17)$$

again whenever $\nu_{m,d} = m/d \rightarrow \infty$, where

$$V_x^{p, \gamma} := \sum_{i=1}^m Y_i^{p, \gamma} \mathbf{I}\{X_i^p = x\}, \quad (2.18)$$

the vote “balance” at x based on the random “sample” $Z_m^{p, \gamma}$ as in (2.3).

Here, as usual, the asymptotic equivalence $A \sim B$ means $A/B \rightarrow 1$.

Display (2.17) shows that the (asymptotically negligible) pairwise differences between (i) the minimax expected excess risk $\inf_{L \in \mathcal{L}} \sup_D \mathfrak{R}_m(L, D)$, (ii) its “randomized” version $\inf_{L \in \mathcal{L}_{\text{rand}}} \sup_D \mathfrak{R}_m(L, D)$, and (iii) the maximum risk $\sup_D \mathfrak{R}_m(L_{m,d;\text{ERM}}, D)$ of any empirical-risk-minimizing learning algorithms of the form L_{ERM} are entirely explained by ties in the mentioned “voting”, when the “no-majority” event $V_x^{p, \gamma} = 0$ occurs for at least one $x \in \mathcal{X} = [d]$.

It is obvious from (2.12) that

$$B(m, d) \geq \sum_{x=1}^d \frac{1}{d} b \mathbf{E} \text{ bayes}(N_x, b) = b \mathbf{E} \text{ bayes}(N_1, b)$$

for any $b \in [0, 1]$, where N_x stands for N_x^p with $p_x = \frac{1}{d}$ for all $x \in [d]$. Thus, in view of (2.16), one immediately obtains

Theorem 2.5.

$$\inf_{L \in \mathcal{L}} \sup_D \mathfrak{R}_m(L, D) \geq \inf_{L \in \mathcal{L}_{\text{rand}}} \sup_D \mathfrak{R}_m(L, D) \geq B_0(m, d) := \sup_{b \in [0, 1]} b \mathbf{E} \text{ bayes}(N, b), \quad (2.19)$$

where N is a binomial r.v. with parameters m and $1/d$.

Let

$$D_\gamma := D_{p,\gamma} \quad \text{and} \quad Z_m^\gamma := Z_m^{p,\gamma} \quad \text{when } p_x = \frac{1}{d} \text{ for all } x \in \mathcal{X} = [d] \quad (2.20)$$

(cf. (2.3)) and

$$\text{ave}_{\gamma \in \{-b,b\}^{[d]}} := \frac{1}{2^d} \sum_{\gamma \in \{-b,b\}^{[d]}} , \quad (2.21)$$

the average over all $\gamma \in \{-b,b\}^{[d]}$, where $b \in [0, 1]$.

Theorem 2.6. *For any $b \in [0, 1]$,*

$$\inf_{L \in \mathcal{L}_{\text{rand}}} \text{ave}_{\gamma \in \{-b,b\}^{[d]}} \mathfrak{R}(L, D_\gamma) = b \, \mathbb{E} \text{ bayes}(N, b). \quad (2.22)$$

As we shall see, Theorem 2.6 follows immediately from the proof of Theorem 2.1. On the other hand, Theorem 2.6 could be viewed as an improvement over Theorem 2.5, because clearly $\text{ave}_{\gamma \in \{-b,b\}^{[d]}} \mathfrak{R}(L, D_\gamma) \leq \sup_D \mathfrak{R}(L, D)$ for any L . Even though the improvement is slight, Theorem 2.6 will be useful, in particular, in the proof of Theorem 1.2.

Remark 2.7. Note that, by (2.13)–(2.14), $\text{bayes}(k, b)$ is a polynomial in b of degree $\leq k$. Hence, $b \, \mathbb{E} \text{ bayes}(N, b)$ is a polynomial in b of degree $\leq m + 1$, and so, the lower bound $B_0(m, d)$ in (2.19) is an algebraic number, which is not hard to compute unless m is too large. For instance, for $c_0(m, d) := B_0(m, d) \sqrt{m/d}$ we find

$$c_0(5, 2) = 0.16757\dots, \quad c_0(50, 20) = 0.17467\dots, \quad \text{and} \quad c_0(50, 2) = 0.16968\dots \quad (2.23)$$

(with the execution times in Mathematica about 0.02 sec, 1.4 sec, and 1 sec, respectively). One may note that, even for such a rather small value $5/2 = 50/20 = 2.5$ of $\nu = m/d$, the values of $c_0(m, d)$ are close to the limit value $c_\infty = 0.16997\dots$ — cf. (1.13) and (2.16). However, more work needs to be done to more fully understand the manner in which the lower bound $B_0(m, d)$ depends on m and d .

The important first step toward this goal is establishing the following convexity property of the function $k \mapsto \text{bayes}(k, b)$.

Proposition 2.8. *Take any $b \in [0, 1]$. Then the largest convex function $[0, \infty) \ni \kappa \mapsto \widetilde{\text{bayes}}(\kappa, b)$ such that $\widetilde{\text{bayes}}(k, b) \leq \text{bayes}(k, b)$ for all $k \in \{0, 1, \dots\}$ is given by the formula*

$$\begin{aligned} & \widetilde{\text{bayes}}(\kappa, b) \\ &:= \begin{cases} (1 - \kappa) \text{bayes}(0, b) + \kappa \text{bayes}(1, b) = \frac{1}{2} (1 - \kappa b) & \text{if } 0 \leq \kappa \leq 1, \\ \frac{2i+3-\kappa}{2} \text{bayes}(2i+1, b) + \frac{\kappa-2i-1}{2} \text{bayes}(2i+3, b) & \text{if } 2i+1 \leq \kappa \leq 2i+3 \end{cases} \end{aligned} \quad (2.24)$$

for any $i \in \overline{0, \infty}$.

That is, the largest convex minorant $\widetilde{\text{bayes}}(\cdot, b)$ on $[0, \infty)$ of the function $\text{bayes}(\cdot, b)$ on $\{0, 1, \dots\}$ is just the linear interpolation of $\text{bayes}(\cdot, b)$ at the points $0, 1, 3, 5, \dots$.

Recall the definition (1.7) of ν . Using (2.19), Proposition 2.8, Jensen's inequality, and the equality $\mathbb{E} N = \nu$, one immediately obtains

Theorem 2.9.

$$\inf_L \sup_D \mathfrak{R}(L, D) \geq B_0(m, d) \geq B_1(\nu) := \sup_{b \in (0, 1)} b \widetilde{\text{bayes}}(\nu, b). \quad (2.25)$$

Here and in the rest of this section, \inf_L can be replaced by either $\inf_{L \in \mathcal{L}}$ or $\inf_{L \in \mathcal{L}_{\text{rand}}}$.

Remark 2.10. An advantage of the lower bound $B_1(\nu)$ in (2.25) over the bound $B_0(m, d)$ in (2.19) is that it depends only on $\nu = m/d$; also, $B_1(\nu)$ is not hard to compute unless ν is too large. Yet, the nature of the dependence of $B_1(\nu)$ on ν may still seem rather obscure. Therefore, we are going to present a lower bound on $B_1(\nu)$ that is much easier to grasp and yet is (i) asymptotic to the original lower bound $B_0(m, d)$ for $\nu = m/d \rightarrow \infty$ and (ii) close to $B_0(m, d)$ even for moderate values of $\nu = m/d$.

In Appendix A, we shall obtain explicit and rather tight lower bounds on the function bayes . In view of Theorem 2.9 and Proposition 2.8, this will result in explicit lower bounds on the minimax excess risk $\inf_L \sup_D \mathfrak{R}(L, D)$, as follows.

Let

$$z_* = 0.75179 \dots \quad (2.26)$$

be the unique maximizer of $\frac{z}{2} (1 - \text{erf}(z/\sqrt{2}))$ in real $z > 0$, with the maximum value $c_\infty = 0.16997 \dots$, as in (1.13).

Theorem 2.11. Assume that $\nu \geq 1$. Let $i_\nu := \lfloor \frac{\nu-1}{2} \rfloor$. Then

$$\inf_L \sup_D \mathfrak{R}(L, D) \geq B_1(\nu) \geq B_2(\nu) := \frac{c_\nu}{\sqrt{\nu}}, \quad (2.27)$$

where

$$c_\nu := \frac{z_*}{2} \left(1 - C_{i_\nu} \frac{\text{erf}(z_*/\sqrt{2})}{1 - z_*^2/(6\nu)} \right) \quad \text{and} \quad C_i = 2^{-2i} \sqrt{\pi(i+1/2)} \binom{2i}{i} \quad (2.28)$$

for $i = 0, 1, \dots$. Moreover, for $\nu \geq 3$, $B_2(\nu)$ admits a simple lower bound on it:

$$B_2(\nu) \geq \tilde{B}_2(\nu) := \frac{\tilde{c}_\nu}{\sqrt{\nu}}, \quad \text{where} \quad \tilde{c}_\nu := \frac{z_*}{2} \left(1 - \left(\frac{i_\nu + 1}{i_\nu} \right)^{1/8} \frac{\text{erf}(z_*/\sqrt{2})}{1 - z_*^2/(6\nu)} \right) \leq c_\nu. \quad (2.29)$$

Remark 2.12. To obtain the second inequality in (2.27) ($B_1(\nu) \geq B_2(\nu)$), in the proof of Theorem 2.11 we are going to use, in particular, two facts: (i) that C_i decreases in i (as stated in Lemma A.2) and (ii) the concavity of $\text{erf}(b\sqrt{k/2})$

in k . If one also uses the obvious fact that $\operatorname{erf}(b\sqrt{k/2})$ increases in k , then, by Chebyshev's integral inequality,

$$\begin{aligned} & (1 - w_i)C_i \operatorname{erf}\left(b\sqrt{\frac{2i+1}{2}}\right) + w_i C_{i+1} \operatorname{erf}\left(b\sqrt{\frac{2i+3}{2}}\right) \\ & \leq [(1 - w_i)C_i + w_i C_{i+1}] \left[(1 - w_i) \operatorname{erf}\left(b\sqrt{\frac{2i+1}{2}}\right) + w_i \operatorname{erf}\left(b\sqrt{\frac{2i+3}{2}}\right) \right] \\ & \leq [(1 - w_i)C_i + w_i C_{i+1}] \operatorname{erf}(b\sqrt{\nu}), \end{aligned}$$

where $i := i_\nu$ and $w_i := \frac{\nu - 2i - 1}{2} \in [0, 1)$. Thus, one can replace $C_{i_\nu} = C_{i_\nu} \vee C_{i_\nu+1}$ in (2.28) by the smaller (and hence better) value $(1 - w_i)C_i + w_i C_{i+1}$, with $i = i_\nu$. Quite similarly, one can replace $\tilde{C}_{i_\nu} := \left(\frac{i_\nu+1}{i_\nu}\right)^{1/8} = \tilde{C}_{i_\nu} \vee \tilde{C}_{i_\nu+1}$ in (2.29) by the smaller (and hence better) value $(1 - w_i)\tilde{C}_i + w_i \tilde{C}_{i+1}$, with $i = i_\nu$. However, these improvements are comparatively small, especially for larger values of ν , and the resulting expressions will be less easy to perceive.

It is clear that

$$c_\nu \rightarrow c_\infty \quad \text{and} \quad \tilde{c}_\nu \rightarrow c_\infty \quad (2.30)$$

as $\nu \rightarrow \infty$. In fact, c_ν and even \tilde{c}_ν are rather close to c_∞ already for rather small values of ν . E.g., one has $c_5 = 0.15532\dots$, $\tilde{c}_5 = 0.15510\dots$, $c_{50} = 0.16852\dots$, and $\tilde{c}_{50} = 0.16852\dots$, and indeed all these four values are rather close to $c_\infty = 0.16997\dots$. We also see that the values of \tilde{c}_ν are not only simpler to compute than, but also very close to, the corresponding values of c_ν .

Inequality (2.27) in Theorem 2.11 does not cover the case $0 < \nu < 1$, and inequality (2.29) does not cover the case $0 \leq \nu < 3$. These two apparently less important cases are covered, complementarily, by

Proposition 2.13.

$$\inf_L \sup_D \mathfrak{R}(L, D) \geq \hat{B}_2(\nu) := \begin{cases} B_1(\nu) = \frac{1}{2}(1 - \nu) & \text{if } 0 < \nu \leq \frac{1}{2}, \\ B_1(\nu) = \frac{1}{8\nu} & \text{if } \frac{1}{2} \leq \nu \leq 1, \\ \frac{(17 - 2\nu)(57187 - 3253\nu - 138\nu^2 + 212\nu^3 - 8\nu^4)}{6480000} & \text{if } 1 \leq \nu \leq 3. \end{cases} \quad (2.31)$$

Remark 2.14. In particular, $\hat{B}_2(1) = B_1(1) = 0.125$, $\hat{B}_2(3) = 0.087018\dots = \frac{0.15072\dots}{\sqrt{3}}$, and $B_1(3) = 0.087019\dots = \frac{0.15072\dots}{\sqrt{3}}$ (cf. (2.27)). More generally, the choices $b = 1$ for $\nu \in (0, \frac{1}{2}]$ and $b = \frac{1}{2\nu}$ for $\nu \in [\frac{1}{2}, 1]$ in the proof of Proposition 2.13 are optimal, in the sense that $\hat{B}_2(\nu) = B_1(\nu)$ for $\nu \in (0, 1]$, as indicated in (2.31). The choice $b = \frac{1}{30}(17 - 2\nu)$ for $\nu \in [1, 3]$ in the just mentioned proof is nearly optimal; namely, then $\hat{B}_2(\nu) > B_1(\nu) - 2 \times 10^{-6}$, for all $\nu \in [1, 3]$; see details on this remark in Section 3, right after the proof of Proposition 2.13. Of course, one can also rather easily give an exact algebraic expression for $B_1(\nu)$ with $\nu \in [1, 3]$; however, that expression (in terms of certain roots of certain polynomials in one variable whose coefficients are polynomials in ν) is complicated and therefore omitted here.

Theorem 2.11, Remark 2.12, relations (2.30), Proposition 2.13, and Remark 2.14 are illustrated in Fig. 1.

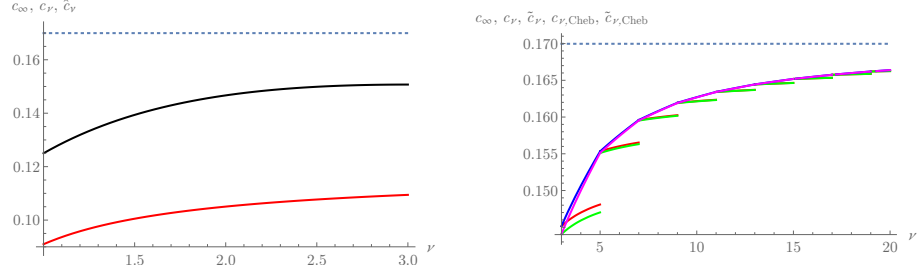


FIG 1. Left panel: graphs of c_ν (red) and $\hat{c}_\nu := \sqrt{\nu} \hat{B}_2(\nu)$ (black) for $\nu \in [1, 3]$. Right panel: graphs of c_ν (red), \tilde{c}_ν (green), $c_{\nu, \text{Cheb}}$ (blue), and $\tilde{c}_{\nu, \text{Cheb}}$ (magenta) for $\nu \in [3, 20]$, where $c_{\nu, \text{Cheb}}$ and $\tilde{c}_{\nu, \text{Cheb}}$ are obtained from the expressions for c_ν and \tilde{c}_ν in (2.28) and (2.29) by replacing there C_{i_ν} and $\tilde{C}_{i_\nu} = \left(\frac{i_\nu+1}{i_\nu}\right)^{1/8}$ by the “Chebyshev” expressions $(1-w_i)C_i + w_i C_{i+1}$ and $(1-w_i)\tilde{C}_i + w_i \tilde{C}_{i+1}$, with $i = i_\nu$ and $w_i := \frac{\nu-2i-1}{2}$, as discussed in Remark 2.12. The dotted horizontal line in both panels is at the level of $c_\infty = 0.16997\dots$.

Let us also present the following very simple, but suboptimal, lower bound — cf. e.g. (2.16).

Proposition 2.15. *If $\nu \geq \frac{3}{41}$, then*

$$\inf_L \sup_D \mathfrak{R}(L, D) \geq B_0(m, d) \geq \frac{0.125}{\sqrt{\nu}}. \quad (2.32)$$

Still, the constant 0.125 in (2.32) is almost twice as good as the mentioned corresponding constant 0.06753 implicit in [2] (see Appendix B for details on the latter value, 0.06753).

Note that the restriction $\nu \geq \frac{3}{41}$ in Proposition 2.15 cannot be dropped, and it is in fact rather close to necessity. Indeed, in view of Remark 2.3, the lower bound $\frac{0.125}{\sqrt{\nu}}$ in (2.32) cannot hold for $\nu < \frac{1}{16} = \frac{3}{48}$.

In conclusion of this section, we summarize the asymptotic behavior of the lower bounds $B_0(m, d), B_1(\nu), B_2(\nu), \tilde{B}_2(\nu)$ on the minimax expected excess risk, as well as the asymptotic behavior of the minimax expected excess risk itself.

Theorem 2.16.

$$\frac{c_\infty}{\sqrt{\nu}} \sim \inf_L \sup_D \mathfrak{R}(L, D) \geq B_0(m, d) \geq B_1(\nu) \geq B_2(\nu) \geq \tilde{B}_2(\nu) \sim \frac{c_\infty}{\sqrt{\nu}} \quad (2.33)$$

as m and d vary in any way such that $\nu = m/d \rightarrow \infty$.

Thus, in view of (2.1), the limit relation in (1.13) holds and, moreover, all the lower bounds $B_0(m, d), B_1(\nu), B_2(\nu), \tilde{B}_2(\nu)$ on the minimax expected excess risk are asymptotically equivalent to the minimax expected excess risk itself whenever $\nu = m/d \rightarrow \infty$. Clearly, Theorem 2.16 complements Theorem 2.4.

3. Proofs

In this section, we shall prove (or provide details for) Theorems 2.1 and 2.6, Proposition 2.8, Theorem 2.11, Proposition 2.13, Remark 2.14, Proposition 2.15, Theorems 2.4 and 2.16 (together), and finally Theorem 1.2, in this order.

Proof of Theorem 2.1. The first equality in (2.12) can be obtained using the von Neumann minimax duality theorem for bilinear functions on the product of simplexes [18] (plus a certain symmetrization argument); more general minimax duality theorems, for convex-concave-like functions, were given in [16], and in [12] a necessary and sufficient condition for the minimax duality for such functions was given.

However, here we are going to offer a more direct and explicit argument, using the explicit form of the to-be-proved minimax decision rule L_{ERM}^* , as defined in (2.10).

To gain some insight, let us begin with the simple case $d = 1$. In that case, $\mathcal{X} = [d] = [1] = \{1\}$, $p = (1)$ (that is, $p_1 = 1$) and $\gamma = (b)$ with $b := \gamma_1 \in [-1, 1]$; also, in the just mentioned definition (2.10) of L_{ERM}^* , the terms x , v_x , i_x , and n_x simplify, respectively, to 1, $v_1 = \sum_{i=1}^m y_i$, 1, and $n_1 = m$, in accordance with the definitions of v_x , i_x , and n_x in (2.7) and (2.11). Here we also have $X_i = 1$ for all i and hence, in view of (2.3), $Z_m^{p,\gamma} = Z_m^b := ((1, Y_1^b), \dots, (1, Y_m^b))$, where $Y_i^b := Y_i^{p,\gamma}$, for $p = (1)$ and $\gamma = (b)$ with $b = \gamma_1$. Let $V_m^b := Y_1^b + \dots + Y_m^b$.

A standard argument along the lines of the proof of the Neyman-Pearson lemma [10] shows that $L_{1,\text{ERM}}^*$ is an optimal Bayesian decision rule, in the sense of being a minimizer of the average $\frac{1}{2} \sum_{y \in \mathcal{Y}} \mathbb{P}(L(Z_m^{by}, U)(1) \neq y)$ of the types

I and II error probabilities over all $L \in \mathcal{L}_{\text{rand},1}$. By symmetry, without loss of generality $b \geq 0$, and so, $b \in [0, 1]$. Then for the corresponding Bayes risk $\frac{1}{2} \sum_{y \in \mathcal{Y}} \mathbb{P}(L_{1,\text{ERM}}^*(Z_m^{by}, U)(1) \neq y)$ for $m \geq 1$ one has

$$\begin{aligned}
 & 2 \frac{1}{2} \sum_{y \in \mathcal{Y}} \mathbb{P}(L_{1,\text{ERM}}^*(Z_m^{by}, U)(1) \neq y) \\
 &= \mathbb{P}(V_m^b < 0) + \mathbb{P}(V_m^b = 0, Y_1^b < 0) + \mathbb{P}(V_m^{-b} > 0) + \mathbb{P}(V_m^{-b} = 0, Y_1^{-b} > 0) \\
 &= \mathbb{P}(V_m^b < 0) + \frac{1}{2} \mathbb{P}(V_m^b = 0) + \mathbb{P}(V_m^{-b} > 0) + \frac{1}{2} \mathbb{P}(V_m^{-b} = 0) \\
 &= \mathbb{P}(V_m^b < 0) + \mathbb{P}(V_m^b = 0) + \mathbb{P}(V_m^{-b} > 0) \\
 &= \mathbb{P}(V_m^b \leq 0) + \mathbb{P}(V_m^{-b} > 0) \\
 &= 1 - (\mathbb{P}(V_m^b > 0) - \mathbb{P}(V_m^{-b} > 0)) = 2 \text{ bayes}(m, b),
 \end{aligned} \tag{3.1}$$

in accordance with (2.13), (2.14), and the assumption $b \in [0, 1]$.

Thus, the Bayes risk $\frac{1}{2} \sum_{y \in \mathcal{Y}} \mathbb{P}(L_{1,\text{ERM}}^*(Z_m^{by}, U)(1) \neq y)$ equals $\text{bayes}(m, b)$ for $m \geq 1$. This conclusion also trivially holds for $m = 0$, in which case both the Bayes risk and $\text{bayes}(m, b)$ equal $\frac{1}{2}$.

Moreover, for all $m \in \overline{0, \infty}$, the Bayes rule $L_{1,\text{ERM}}^*$ is a risk equalizer, in the sense that $\mathbb{P}(L_{1,\text{ERM}}^*(Z_m^{by}, U)(1) \neq y)$ does not depend on the choice of $y \in \mathcal{Y} =$

$\{-1, 1\}$; this follows because (i) $(L_{1,\text{ERM}}^*((Z_m^b)^-, -U) = -L_{1,\text{ERM}}^*(Z_m^b, U)$, where $(Z_m^b)^- := ((1, -Y_1^b), \dots, (1, -Y_m^b))$, and (ii) the distribution of $(Y_1^{-b}, \dots, Y_k^{-b}, -U)$ is the same as that of $-(Y_1^b, \dots, Y_k^b, U)$.

Let us now proceed to the general case of any natural d , which will in a sense be reduced to the case $d = 1$. Take any $m \in \overline{0, \infty}$, any randomized learning algorithm $L: (\mathcal{X} \times \mathcal{Y})^m \times [-1, 1] \rightarrow \mathcal{H}$, any $p \in [0, 1]^{[d]}$ such that $\sum_{x=1}^d p_x = 1$, and any $\gamma \in [-1, 1]^{[d]}$. For each $x \in [d]$, introduce the random set

$$\mathcal{J}_x^p := \{i \in [m]: X_i^p = x\}$$

and its cardinality

$$N_x^p := \text{card } \mathcal{J}_x^p.$$

Then, by (2.9),

$$\mathfrak{R}_m(L; p, \gamma) = \sum_{x=1}^d p_x |\gamma_x| \sum_{k=0}^m \mathbb{P}(L(Z_m^{p,\gamma}, U)(x) \neq h_\gamma(x), N_x^p = k). \quad (3.2)$$

Next, take any $x \in [d]$ and any $k = 0, \dots, m$. Then

$$\begin{aligned} \mathbb{P}(L(Z_m^{p,\gamma}, U)(x) \neq h_\gamma(x), N_x^p = k) \\ = \sum_{J \in \binom{[m]}{k}} \mathbb{P}(L(Z_m^{p,\gamma}, U)(x) \neq h_\gamma(x), \mathcal{J}_x^p = J), \end{aligned} \quad (3.3)$$

where $\binom{[m]}{k} := \{J \subseteq [m]: \text{card } J = k\}$.

Further, take any set $J \in \binom{[m]}{k}$. Writing J as $\{i_1, \dots, i_k\}$ with $i_1 < \dots < i_k$, let $X_J^p := (X_{i_1}^p, \dots, X_{i_k}^p)$, and similarly define $X_{J^c}^p, Y_J^{p,\gamma}$, and $Y_{J^c}^{p,\gamma}$, where $J^c := [m] \setminus J$. Then

$$L(Z_m^{p,\gamma}, U)(x) = H_{L,x,J}(X_J^p, Y_J^{p,\gamma}, X_{J^c}^p, Y_{J^c}^{p,\gamma}, U) \quad (3.4)$$

for some Borel-measurable function

$$H_{L,x,J}: \mathcal{X}^k \times \mathcal{Y}^k \times \mathcal{X}^{m-k} \times \mathcal{Y}^{m-k} \times [-1, 1] \rightarrow \{-1, 1\}.$$

Let $x^J := (x, \dots, x) \in \mathcal{X}^k$. Then

$$\begin{aligned} \mathbb{P}(L(Z_m^{p,\gamma}, U)(x) \neq h_\gamma(x), \mathcal{J}_x^p = J) \\ = \sum_{\xi \in (\mathcal{X} \setminus \{x\})^{m-k}} \sum_{\eta \in \mathcal{Y}^{m-k}} P'(x, J, \xi, \eta, p, \gamma_{[x]}) P''(L, x, J, \xi, \eta, |\gamma_x|; \text{sgn } \gamma_x), \end{aligned} \quad (3.5)$$

where $P'(x, J, \xi, \eta, p, \gamma_{[x]}) := \mathbb{P}(X_J^p = x^J, X_{J^c}^p = \xi, Y_{J^c}^{p,\gamma} = \eta)$, $\gamma_{[x]}$ denotes the restriction of the function γ to the set $\mathcal{X} \setminus \{x\}$, and

$$\begin{aligned} P''(L, x, J, \xi, \eta, p, |\gamma_x|; \sigma_x) \\ := \mathbb{P}(H_{L,x,J}(x^J, Y_J^{p,\gamma}, \xi, \eta, U) \neq \sigma_x | X_J^p = x^J, X_{J^c}^p = \xi, Y_{J^c}^{p,\gamma} = \eta) \end{aligned}$$

for $\sigma_x \in \{-1, 1\}$. That the notation for the arguments of the functions P' and P'' — as far as the dependence of the values of these functions on γ is concerned — is justified will be clear in a moment.

At this point, note that, by (3.2), (3.3), and (3.5),

$$\mathfrak{R}_m(L; p, \gamma) = \sum_{x, k, J, \xi, \eta} p_x |\gamma_x| P'(x, J, \xi, \eta, p, \gamma_{[x]}) P''(L, x, J, \xi, \eta, |\gamma_x|; \text{sgn } \gamma_x), \quad (3.6)$$

where $\sum_{x, k, J, \xi, \eta} := \sum_{x=1}^d \sum_{k=0}^m \sum_{J \in \binom{[m]}{k}} \sum_{\xi \in (\mathcal{X} \setminus \{x\})^{m-k}} \sum_{\eta \in \mathcal{Y}^{m-k}}$, and, moreover, the sum

$$\sum_{J, \xi, \eta} P'(x, J, \xi, \eta, p, \gamma_{[x]}) = \mathbb{P}(N_x^p = k) \quad (3.7)$$

(where $\sum_{J, \xi, \eta} := \sum_{J \in \binom{[m]}{k}} \sum_{\xi \in (\mathcal{X} \setminus \{x\})^{m-k}} \sum_{\eta \in \mathcal{Y}^{m-k}}$) does not depend on γ . To quickly see why identity (3.7) holds, look back at (3.3) and (3.5), with the event $\{L(Z_m^{p, \gamma}, U)(x) \neq h_\gamma(x)\}$ replaced there by an event of probability 1.

Because of the independence of $(X_1^p, Y_1^{p, \gamma}), \dots, (X_m^p, Y_m^{p, \gamma}), U$, one can observe that for any $L \in \mathcal{L}_{\text{rand}}$, $x \in [d]$, $k \in \overline{0, m}$, $J \in \binom{[m]}{k}$, $\xi \in (\mathcal{X} \setminus \{x\})^{m-k}$, $\eta \in \mathcal{Y}^{m-k}$, and $p \in [0, 1]^{[d]}$ such that $\sum_{x=1}^d p_x = 1$, the conditional probability $P''(L, x, J, \xi, \eta, |\gamma_x|; \text{sgn } \gamma_x)$ in (3.5) depends on γ only through γ_x , whereas the unconditional probability $P'(x, J, \xi, \eta, p, \gamma_{[x]})$ in (3.5) depends on γ only through $\gamma_{[x]} = \gamma|_{\mathcal{X} \setminus \{x\}}$. In particular, this observation justifies the notation for the arguments of the functions P' and P'' .

More importantly, introducing the averaging operators

$$\text{ave}_\sigma := \frac{1}{2^d} \sum_{\sigma \in \{-1, 1\}^{[d]}} \quad , \quad \text{ave}_{\sigma_{[x]}} := \frac{1}{2^{d-1}} \sum_{\sigma_{[x]} \in \{-1, 1\}^{\mathcal{X} \setminus \{x\}}} \quad , \quad \text{ave}_{\sigma_x} := \frac{1}{2} \sum_{\sigma_x \in \{-1, 1\}} \quad ,$$

in view of (3.6) one has

$$\begin{aligned} \text{ave}_\sigma \mathfrak{R}_m(L; p, |\gamma| \sigma) &= \sum_{x, k, J, \xi, \eta} p_x |\gamma_x| \text{ave}_{\sigma_{[x]}} P'(x, J, \xi, \eta, p, |\gamma_{[x]}| \sigma_{[x]}) \\ &\quad \times \text{ave}_{\sigma_x} P''(L, x, J, \xi, \eta, |\gamma_x|; \sigma_x). \end{aligned} \quad (3.8)$$

Next, for any $L \in \mathcal{L}_{\text{rand}}$, $x \in [d]$, $k \in \overline{0, m}$, $J \in \binom{[m]}{k}$, $\xi \in (\mathcal{X} \setminus \{x\})^{m-k}$, $\eta \in \mathcal{Y}^{m-k}$, and $p \in [0, 1]^{[d]}$ such that $\sum_{x=1}^d p_x = 1$, the conditional distribution of $(Y_J^{p, \gamma}, \xi, \eta, U)$ given $X_J^p = x^J, X_{J^c}^p = \xi, Y_{J^c}^{p, \gamma} = \eta$ is the same as that of $(Y_1^{\gamma_x}, \dots, Y_k^{\gamma_x}, U)$, where the $Y_i^{\gamma_x}$'s are the same as in the consideration of the case $d = 1$ in the beginning of this proof. Therefore, in view of (3.1) and (2.10),

$$\text{ave}_{\sigma_x} P''(L, x, J, \xi, \eta, |\gamma_x|; \sigma_x) \geq \text{ave}_{\sigma_x} P''(L_{m, \text{ERM}}^*, x, J, \xi, \eta, |\gamma_x|; \sigma_x) = \text{bayes}(k, |\gamma_x|).$$

So, by (3.8) and because $P'(x, J, \xi, \eta, p, |\gamma_{[x]}| \sigma_{[x]})$ does not depend on L , one has

$$\text{ave}_\sigma \mathfrak{R}_m(L; p, |\gamma| \sigma) \geq \text{ave}_\sigma \mathfrak{R}_m(L_{m, \text{ERM}}^*; p, |\gamma| \sigma). \quad (3.9)$$

Moreover, $P''(L_{m, \text{ERM}}^*, x, J, \xi, \eta, |\gamma_x|; \sigma_x) = \text{bayes}(k, |\gamma_x|)$ does not depend on the choice of $\sigma_x \in \{-1, 1\}$, and so, by (3.6) and (3.7), for any $\gamma \in [-1, 1]^d$

$$\begin{aligned} \mathfrak{R}_m(L_{m, \text{ERM}}^*; p, \gamma) &= \sum_{x, k, J, \xi, \eta} p_x |\gamma_x| P'(x, J, \xi, \eta, p, \gamma_{[x]}) \text{bayes}(k, |\gamma_x|) \\ &= \sum_{x=1}^d p_x |\gamma_x| \sum_{k=0}^m \text{bayes}(k, |\gamma_x|) \sum_{J, \xi, \eta} P'(x, J, \xi, \eta, p, |\gamma_{[x]}| \sigma_{[x]}) \\ &= \sum_{x=1}^d p_x |\gamma_x| \sum_{k=0}^m \text{bayes}(k, |\gamma_x|) \mathbf{P}(N_x^p = k) \\ &= \sum_{x=1}^d p_x |\gamma_x| \mathbf{E} \text{bayes}(N_x^p, |\gamma_x|), \end{aligned} \quad (3.10)$$

which proves (2.15). This and (3.9) yield

$$\begin{aligned} \max_{\sigma \in \{-1, 1\}^d} \mathfrak{R}_m(L_{m, \text{ERM}}^*; p, |\gamma| \sigma) &= \text{ave}_\sigma \mathfrak{R}_m(L_{m, \text{ERM}}^*; p, |\gamma| \sigma) \\ &= \sum_{x=1}^d p_x |\gamma_x| \mathbf{E} \text{bayes}(N_x^p, |\gamma_x|) \leq \text{ave}_\sigma \mathfrak{R}_m(L; p, |\gamma| \sigma) \leq \max_{\sigma \in \{-1, 1\}^d} \mathfrak{R}_m(L; p, |\gamma| \sigma). \end{aligned}$$

Taking now $\sup_{p, \gamma}$, one completes the proof of Theorem 2.1. \square

Proof of Theorem 2.6. This theorem follows immediately from (3.9) and (3.10) by taking there $p_x = \frac{1}{d}$ for all $x \in [d]$ and any $\gamma \in \{-b, b\}^{[d]}$. \square

Proof of Proposition 2.8. By (2.13)–(2.14) and the law of large numbers, $\text{bayes}(k, b) \rightarrow 0$ as $k \rightarrow \infty$. Also, $\text{bayes}(0, b) = \frac{1}{2}$. So, by Lemma A.1 in Appendix A, $\text{bayes}(k, b)$ is convex and decreasing in $k \in \{0, 1, 3, 5, \dots\}$. It remains to use relations (A.9) and (A.7) in Appendix A and, again, (2.13). \square

Proof of Theorem 2.11. The first inequality in (2.27) comes from (2.25). The second inequality in (2.27) follows immediately from Lemma A.3 with $\kappa = \nu$ and $b = z_*/\sqrt{\nu}$. The inequalities in (2.29) follow immediately from (2.27), (2.28), and (A.13). \square

Proof of Proposition 2.13. By (2.24), (2.13), and (A.9),

$$\widetilde{\text{bayes}}(\nu) = \begin{cases} \frac{1}{2}(1 - \nu b) & \text{if } 0 < \nu \leq 1, \\ \frac{1}{8}(4 + b^3(\nu - 1) - b(3 + \nu)) & \text{if } 1 \leq \nu \leq 3. \end{cases}$$

Recalling now (2.25) and using the values $b = 1$ for $\nu \in (0, \frac{1}{2}]$, $b = \frac{1}{2\nu}$ for $\nu \in [\frac{1}{2}, 1]$, and $b = \frac{1}{30}(17 - 2\nu)$ for $\nu \in [1, 3]$, one obtains (2.31). \square

Details on Remark 2.14. The inequality $\hat{B}_2(\nu) > B_1(\nu) - 2 \times 10^{-6}$ for all $\nu \in [1, 3]$, mentioned in that remark, can be verified, e.g., by issuing the Mathematica command `Reduce[b 1/8 (4 + b^3 (nu - 1) - b (3 + nu)) - hB2[nu] >= 2 10^(-6) && 1 <= nu <= 3 && 0 <= b <= 1]`, where `hB2[nu]` stands for $\hat{B}_2(\nu)$. This command then outputs `False`, which means that indeed $B_1(\nu) = \max_{0 \leq b \leq 1} \widehat{b \text{ bayes}}(\nu) = \max_{0 \leq b \leq 1} b \frac{1}{8} (4 + b^3(\nu - 1) - b(3 + \nu)) < \hat{B}_2(\nu) + 2 \times 10^{-6}$. \square

Proof of Proposition 2.15. The first inequality in (2.32) holds by (2.19). Take now any $b \in (0, 1]$. From (A.5), Lemma A.2, and inequality $S_q(b) \leq b$, it follows that

$$s_k(b) \leq b\sqrt{k} \quad (3.11)$$

for odd natural k . By (A.7), inequality (3.11) holds for even natural k as well, and it trivially holds for $k = 0$. Using now the definition of $B_0(m, d)$ in (2.19) together with (2.13), (3.11), and Jensen's inequality, noticing that $\frac{1}{2\sqrt{\nu}} \in (0, 1]$ if $\nu \geq \frac{1}{4}$, and substituting $\frac{1}{2\sqrt{\nu}}$ for b , one has

$$\begin{aligned} B_0(m, d) &\geq b \mathbb{E} \text{ bayes}(N, b) \geq \frac{b}{2} (1 - b \mathbb{E} \sqrt{N}) \geq \frac{b}{2} (1 - b\sqrt{\mathbb{E} N}) \\ &= \frac{b}{2} (1 - b\sqrt{\nu}) = \frac{0.125}{\sqrt{\nu}}, \end{aligned}$$

in the case when $\nu \geq \frac{1}{4}$.

It remains to consider the case when $\frac{1}{4} > \nu \geq \frac{3}{41}$. Then, by (2.25) and (2.31),

$$B_0(m, d) \geq B_1(\nu) = \frac{1}{2} (1 - \nu) > \frac{0.125}{\sqrt{\nu}},$$

which completes the proof of Proposition 2.15. \square

Proof of Theorems 2.4 and 2.16. The two inequalities in (2.16) are trivial. The first, second, third, and fourth inequalities in (2.33) were already established as the inequalities in (2.19), the second inequality in (2.25), the second inequality in (2.27), and the first inequality in (2.29), respectively. The second asymptotic equivalence in (2.33) follows immediately from (2.29) and (2.30).

So, in view of (2.12), it suffices to show that (2.17) holds and

$$B(m, d) \stackrel{?}{\lesssim} \frac{c_\infty}{\sqrt{\nu}}, \quad (3.12)$$

where, as usual, $A \lesssim B$ means $\limsup A/B \leq 1$.

In this proof, all the limit relations are stated for $\nu = \nu_{m,d} = m/d \rightarrow \infty$ and all the other relations are stated under the condition that ν is large enough.

By (2.9), (2.6), and (2.18),

$$\mathfrak{R}(L_{\text{ERM}}; p, \gamma) \leq \sum_{x=1}^d p_x |\gamma_x| \left[\mathbb{P}(V_x^{p,\gamma} \neq 0, \text{sgn } V_x^{p,\gamma} \neq \text{sgn } \gamma_x) + \mathbb{P}(V_x^{p,\gamma} = 0) \right]. \quad (3.13)$$

For $b \in [-1, 1]$, $k \in \overline{0, \infty}$, and V_k^b as defined in Theorem 2.1, introduce next

$$Q_{\pm}(k, b) := \mathbb{P}(\pm V_k^b > 0) \quad \text{and} \quad Q_0(k, b) := \mathbb{P}(V_k^b = 0),$$

so that, by (3.1), $\text{bayes}(k, b) = Q_{-}(k, b) + \frac{1}{2} Q_0(k, b) = Q_{+}(k, -b) + \frac{1}{2} Q_0(k, b)$.

If $\gamma_x > 0$ for some $x \in [d]$, then

$$\begin{aligned} \mathbb{P}(V_x^{p, \gamma} \neq 0, \text{sgn } V_x^{p, \gamma} \neq \text{sgn } \gamma_x) &= \mathbb{P}(V_x^{p, \gamma} < 0) = \sum_{k=0}^m \mathbb{P}(N_x^p = k) Q_{-}(k, |\gamma_x|) \\ &= \mathbb{E} Q_{-}(N_x^p, |\gamma_x|) = \mathbb{E} \text{bayes}(N_x^p, |\gamma_x|) - \frac{1}{2} \mathbb{E} Q_0(N_x^p, |\gamma_x|), \end{aligned}$$

where $N_x^p = \sum_{i=1}^m \mathbb{I}\{X_i^p = x\}$, as before. Similarly, the latter expression, $\mathbb{E} \text{bayes}(N_x^p, |\gamma_x|) - \frac{1}{2} \mathbb{E} Q_0(N_x^p, |\gamma_x|)$, for $\mathbb{P}(V_x^{p, \gamma} \neq 0, \text{sgn } V_x^{p, \gamma} \neq \text{sgn } \gamma_x)$ holds when $\gamma_x < 0$ as well. On the other hand, it is similarly seen that

$$\mathbb{E} Q_0(N_x^p, |\gamma_x|) = \mathbb{P}(V_x^{p, \gamma} = 0).$$

Thus, (3.13) yields

$$\mathfrak{R}(L_{\text{ERM}}; p, \gamma) \leq \sum_{x=1}^d p_x |\gamma_x| \mathbb{E} \text{bayes}(N_x^p, |\gamma_x|) + \frac{1}{2} \sum_{x=1}^d p_x |\gamma_x| \mathbb{E} Q_0(N_x^p, |\gamma_x|).$$

In particular, in view of (2.12), this implies the second inequality in (2.17); the first inequality there is trivial.

Now, to complete the proof of (2.17) and Theorem 2.16, it remains to verify (3.12) and

$$\sup_{p, \gamma} \sum_{x=1}^d p_x |\gamma_x| \mathbb{E} Q_0(N_x^p, |\gamma_x|) \stackrel{?}{=} O(1/\nu). \quad (3.14)$$

Take any $b \in (0, 1]$ and any natural $k \geq 3$, so that, by (A.3), $q := q_k \geq 1 > 0$. Note that $s_k(1) = 1$. Hence, by (2.13), (A.5),

$$\begin{aligned} \text{bayes}(k, b) &= \frac{1}{2} (1 - s_k(b)) = \frac{1}{2} (s_k(1) - s_k(b)) = \frac{1}{2} s'_k(0) (S_q(1) - S_q(b)) \\ &= \frac{1}{2} s'_k(0) \int_b^1 (1 - u^2)^q du \leq \frac{1}{2} s'_k(0) \int_b^1 e^{-qu^2} du \leq \frac{1}{2} s'_k(0) \int_b^\infty e^{-qu^2} du \\ &= A_k (1 - \text{erf}(b\sqrt{q})), \end{aligned} \quad (3.15)$$

where $A_k := \frac{s'_k(0)\sqrt{\pi}}{4\sqrt{q}} \rightarrow \frac{1}{2}$ as $k \rightarrow \infty$, by Lemma A.2 and (A.6). Therefore, for $z := b\sqrt{2q}$ one has

$$b \text{bayes}(k, b) \leq \frac{\lambda_k}{\sqrt{k}} \frac{z}{2} (1 - \text{erf}(z/\sqrt{2})) \leq c_\infty \frac{\lambda_k}{\sqrt{k}} \quad (3.16)$$

by (1.13), where

$$\lambda_k := 2A_k \sqrt{\frac{k}{2q}} \rightarrow 1 \quad (3.17)$$

as $k \rightarrow \infty$.

Take now any $\varepsilon \in (0, 1)$. In view of (3.16) and because $\text{bayes} \leq \frac{1}{2}$,

$$b \text{ bayes}(k, b) \leq \frac{A}{\sqrt{k+1}} \quad (3.18)$$

for some real $A > 0$, all $b \in [0, 1]$, and all $k = 0, 1, \dots$.

Since the r.v. N_x^p has the binomial distribution with parameters m and p_x , one has

$$\mathbb{P}(N_x^p \leq (1 - \varepsilon)mp_x) \leq e^{-\varepsilon^2 mp_x/2}.$$

Such an inequality (sometimes attributed to Chernoff) is ubiquitous in computer science (see e.g. [1, Proposition 2.4(a)]), under the names *multiplicative Chernoff* or *Angluin-Valiant* bound. However, we have been unable to find a proof of it in the literature, except that this inequality immediately follows from the more general and precise results in [11, (1.3) or (2.31)] or [13, Theorem 7].

Therefore,

$$\begin{aligned} S_{01} &:= \sum_{x=1}^d p_x \mathbb{E} \frac{1}{\sqrt{N_x^p + 1}} \mathbb{I}\{N_x^p \leq (1 - \varepsilon)mp_x\} \\ &\leq \sum_{x=1}^d p_x \mathbb{P}(N_x^p \leq (1 - \varepsilon)mp_x) \leq \sum_{x=1}^d p_x e^{-\varepsilon^2 mp_x/2} \leq \sum_{x=1}^d a_\varepsilon \frac{1}{m} = a_\varepsilon \frac{d}{m}, \end{aligned} \quad (3.19)$$

where a_ε is a positive real number depending only on ε . Also,

$$\begin{aligned} S_{02} &:= \sum_{x=1}^d p_x \mathbb{E} \frac{1}{\sqrt{N_x^p + 1}} \mathbb{I}\{N_x^p > (1 - \varepsilon)mp_x\} \\ &\leq \sum_{x=1}^d p_x \frac{1}{\sqrt{(1 - \varepsilon)mp_x + 1}} \leq \sum_{x=1}^d \frac{\sqrt{p_x}}{\sqrt{(1 - \varepsilon)m}} \leq \frac{1}{\sqrt{1 - \varepsilon}} \sqrt{\frac{d}{m}}; \end{aligned} \quad (3.20)$$

the last inequality here is obtained using the concavity of the square root function together with the condition $\sum_{x=1}^d p_x = 1$. Thus, by (3.19) and (3.20),

$$\sum_{x=1}^d p_x \mathbb{E} \frac{1}{\sqrt{N_x^p + 1}} = S_{01} + S_{02} \leq \frac{1}{1 - \varepsilon} \sqrt{\frac{d}{m}} \quad (3.21)$$

if m/d is large enough, depending on ε .

In view of (3.18),

$$\begin{aligned}
S_{11} &:= \sum_{x=1}^d p_x |\gamma_x| \mathbb{E} \text{ bayes}(N_x^p, |\gamma_x|) \mathbb{I}\{N_x^p > (1-\varepsilon)mp_x\} \mathbb{I}\{p_x \leq \varepsilon/d\} \\
&\leq \sum_{x=1}^d p_x \mathbb{E} \frac{A}{\sqrt{N_x^p + 1}} \mathbb{I}\{N_x^p > (1-\varepsilon)mp_x\} \mathbb{I}\{p_x \leq \varepsilon/d\} \\
&\leq \sum_{x=1}^d p_x \frac{A}{\sqrt{(1-\varepsilon)mp_x + 1}} \mathbb{I}\{p_x \leq \varepsilon/d\} \leq A \sum_{x=1}^d \frac{\sqrt{p_x}}{\sqrt{(1-\varepsilon)m}} \mathbb{I}\{p_x \leq \varepsilon/d\} \\
&\leq A \sum_{x=1}^d \frac{\sqrt{\varepsilon/d}}{\sqrt{(1-\varepsilon)m}} = A \sqrt{\frac{\varepsilon}{1-\varepsilon}} \sqrt{\frac{d}{m}}.
\end{aligned}$$

Next, taking into account (3.16), (3.17), and (3.20), one has

$$\begin{aligned}
S_{12} &:= \sum_{x=1}^d p_x |\gamma_x| \mathbb{E} \text{ bayes}(N_x^p, |\gamma_x|) \mathbb{I}\{N_x^p > (1-\varepsilon)mp_x\} \mathbb{I}\{p_x > \varepsilon/d\} \\
&\leq \sum_{x=1}^d p_x |\gamma_x| \mathbb{E} \text{ bayes}(N_x^p, |\gamma_x|) \mathbb{I}\{N_x^p > (1-\varepsilon)mp_x\} \mathbb{I}\{N_x^p > (1-\varepsilon)\varepsilon m/d\} \\
&\leq \sum_{x=1}^d p_x (1+\varepsilon) c_\infty \mathbb{E} \frac{1}{\sqrt{N_x^p + 1}} \mathbb{I}\{N_x^p > (1-\varepsilon)mp_x\} = (1+\varepsilon) c_\infty S_{02} \\
&\leq \frac{(1+\varepsilon) c_\infty}{\sqrt{1-\varepsilon}} \sqrt{\frac{d}{m}}.
\end{aligned}$$

So,

$$\begin{aligned}
S_1 &:= \sum_{x=1}^d p_x |\gamma_x| \mathbb{E} \text{ bayes}(N_x^p, |\gamma_x|) \mathbb{I}\{N_x^p > (1-\varepsilon)mp_x\} = S_{11} + S_{12} \\
&\leq \frac{(1 + A_1 \sqrt{\varepsilon}) c_\infty}{\sqrt{1-\varepsilon}} \sqrt{\frac{d}{m}} \quad (3.22)
\end{aligned}$$

for some universal real constant $A_1 > 0$.

On the other hand, by (3.18) and (3.19),

$$\begin{aligned}
S_2 &:= \sum_{x=1}^d p_x |\gamma_x| \mathbb{E} \text{ bayes}(N_x^p, |\gamma_x|) \mathbb{I}\{N_x^p \leq (1-\varepsilon)mp_x\} \\
&\leq AS_{01} \leq Aa_\varepsilon \frac{d}{m} \leq \varepsilon \sqrt{\frac{d}{m}}.
\end{aligned}$$

Combining this with (3.22), we conclude that

$$\sum_{x=1}^d p_x |\gamma_x| \mathbb{E} \text{ bayes}(N_x^p, |\gamma_x|) = S_1 + S_2 \leq \frac{(1 + A_2 \sqrt{\varepsilon}) c_\infty}{\sqrt{1-\varepsilon}} \sqrt{\frac{d}{m}} \quad (3.23)$$

for some universal real constant $A_2 > 0$. Thus, in view of the definition of $B(m, d)$ in (2.12), the asymptotic relation (3.12) is proved.

To complete the proof of Theorems 2.4 and 2.16, let us finally verify (3.14). If $k = 2j$ is even, then

$$bQ_0(k, b) = \binom{2j}{j} \frac{1}{4^j} b(1 - b^2)^j \leq \frac{A_3}{\sqrt{k+1}} b e^{-b^2 k/2} \leq \frac{A_4}{k+1}$$

for some universal real constants $A_3 > 0$ and $A_4 > 0$ and all $b \in [0, 1]$; since $Q_0(k, b) = 0$ if k is odd, the above bound in fact holds for all $k = 0, 1, \dots$. So,

$$\sum_{x=1}^d p_x |\gamma_x| \mathbb{E} Q_0(N_x^p, |\gamma_x|) \leq A_4 \sum_{x=1}^d p_x \mathbb{E} \frac{1}{N_x^p + 1} \leq A_4 \left(a_\varepsilon + \frac{1}{1 - \varepsilon} \right) \frac{d}{m};$$

the second inequality in the above display is obtained similarly to the inequality in (3.21). Thus, (3.14) is verified, and the proof of Theorems 2.4 and 2.16 is complete. \square

Proof of Theorem 1.2. Take any learning algorithm $L \in \mathcal{L}_{\text{rand}}$. Take any $b \in (0, 1]$ and then any $\varepsilon \in (0, b)$ and any $\gamma \in \{-b, b\}^{[d]}$. Let D_γ and Z_m^γ be as in (2.20). Let $\hat{\Delta}^\gamma := \Delta(L(Z_m^\gamma, U), D^\gamma)$. By (2.5), $\hat{\Delta}^\gamma \leq b$. So, $\mathbb{I}\{\hat{\Delta}^\gamma > \varepsilon\} \geq \frac{1}{b-\varepsilon} (\hat{\Delta}^\gamma - \varepsilon)$, whence $\mathbb{P}(\hat{\Delta}^\gamma > \varepsilon) \geq \frac{1}{b-\varepsilon} (\mathbb{E} \hat{\Delta}^\gamma - \varepsilon)$. Therefore, by (2.21), Theorem 2.6, Proposition 2.8, and Jensen's inequality,

$$\begin{aligned} \max_{\gamma \in \{-1, 1\}^{[d]}} \mathbb{P}(\hat{\Delta}^\gamma > \varepsilon) &\geq \text{ave}_{\gamma \in \{-1, 1\}^{[d]}} \mathbb{P}(\hat{\Delta}^\gamma > \varepsilon) \geq \frac{1}{b - \varepsilon} \left(\text{ave}_{\gamma \in \{-1, 1\}^{[d]}} \mathbb{E} \hat{\Delta}^\gamma - \varepsilon \right) \\ &\geq \frac{1}{b - \varepsilon} (b \widehat{\text{bayes}}(\nu, b) - \varepsilon). \end{aligned}$$

Take now any $\nu_* \in [3, \infty)$ and any real $\nu \geq \nu_*$. Then, by Lemma A.3 and (A.13), $\widehat{\text{bayes}}(\nu, b) \geq \frac{1}{2} (1 - (\frac{i_{\nu_*} + 1}{i_{\nu_*}})^{1/8} \psi_b(\nu))$. Further, take any $z \in (0, \sqrt{\nu_*}]$ and $w \in (0, z)$, and then take $b = z/\sqrt{\nu}$ and $\varepsilon = w/\sqrt{\nu}$, so that the conditions $b \in (0, 1]$ and $\varepsilon \in (0, b)$ assumed in the beginning of this proof do hold. It follows that

$$\begin{aligned} \max_{\gamma \in \{-1, 1\}^{[d]}} \mathbb{P} \left(\hat{\Delta}^\gamma > \frac{w}{\sqrt{\nu}} \right) &\geq \frac{1}{z - w} \left(\frac{z}{2} \left[1 - \left(\frac{i_{\nu_*} + 1}{i_{\nu_*}} \right)^{1/8} \frac{\text{erf}(z/\sqrt{2})}{1 - z^2/(6\nu_*)} \right] - w \right) \\ &=: P_{\text{low}}(w, \nu_*, z). \end{aligned}$$

It remains to note that $P_{\text{low}}(\frac{1}{\sqrt{320}}, \frac{128}{10}, \frac{331}{1000}) > 0.238$, $P_{\text{low}}(\frac{1}{\sqrt{320}}, 3, \frac{320}{1000}) > 0.227$, $P_{\text{low}}(\frac{1}{\sqrt{413/10}}, \frac{128}{10}, \frac{681}{1000}) > 0.01563 > \frac{1}{64}$, and $P_{\text{low}}(\frac{1}{\sqrt{496/10}}, 3, \frac{601}{1000}) > 0.0159 > \frac{1}{64}$. \square

Appendix A: Identities and inequalities for binomial distributions: details concerning the function bayes

Recall the definition of bayes in (2.13). Take any $b \in [0, 1]$ and $k \in \overline{1, \infty}$. By (2.14),

$$s_k(b) = \frac{1}{2^k} \sum_{i=0}^j \binom{k}{i} [(1+b)^{k-i}(1-b)^i - (1-b)^{k-i}(1+b)^i], \quad (\text{A.1})$$

$$j := j_k := \lfloor k/2 \rfloor. \quad (\text{A.2})$$

Using identities $(k-i)\binom{k}{i} = k\binom{k-1}{i}$ and $i\binom{k}{i} = k\binom{k-1}{i-1}$, we have

$$\begin{aligned} \frac{2^k}{k} s'_k(b) &:= \sum_{i=0}^j \binom{k-1}{i} [(1+b)^{k-i-1}(1-b)^i + (1-b)^{k-i-1}(1+b)^i] \\ &\quad - \sum_{i=1}^j \binom{k-1}{i-1} [(1+b)^{k-i}(1-b)^{i-1} + (1-b)^{k-i}(1+b)^{i-1}]. \end{aligned}$$

Making in the second sum the substitution $i = r + 1$, then replacing there r back by i , and introducing

$$q := q_k := k - j - 1, \quad (\text{A.3})$$

we have

$$\frac{2^k}{k} s'_k(b) / \binom{k-1}{j} = (1+b)^q(1-b)^j + (1-b)^q(1+b)^j = 2(1-b^2)^q, \quad (\text{A.4})$$

which is non-increasing in $b \in [0, 1]$. So, the function s_k is concave.

Moreover, it follows that

$$s_k(b) = s'_k(0)S_q(b), \quad \text{where} \quad S_q(b) := \int_0^b (1-u^2)^q du. \quad (\text{A.5})$$

For all $i \in \overline{0, \infty}$, in view of (A.2), (A.3), and (A.4),

$$q_{2i+2} = q_{2i+1} = i \quad \text{and} \quad s'_{2i+2}(0) = s'_{2i+1}(0) \quad (\text{A.6})$$

and hence, by (A.5), one has the curious, and useful, identity

$$s_{2i+2}(b) = s_{2i+1}(b). \quad (\text{A.7})$$

We also have

Lemma A.1. *Take any $b \in (0, 1)$. Then the function*

$$\{0, 1, 3, 5, \dots\} \ni k \mapsto \text{bayes}(k, b) \quad (\text{A.8})$$

is strictly convex.

Proof. In view of (2.13), it is enough to show that the function $\{0, 1, 3, 5, \dots\} \ni k \mapsto s_k(b)$ is strictly concave. By (A.1) and (A.2),

$$s_0(b) = 0, \quad s_1(b) = b, \quad s_3(b) = \frac{1}{2}(3b - b^3) < 3s_1(b). \quad (\text{A.9})$$

So, the restriction of the function in (A.8) to the set $\{0, 1, 3\}$ is strictly concave.

It remains to show that the restriction of this function to the set $\{1, 3, 5, \dots\}$ is strictly concave. Take any $i \in \overline{0, \infty}$. We have to show that

$$g(b) := \tilde{s}_i(b) + \tilde{s}_{i+2}(b) - 2\tilde{s}_{i+1}(b) < 0,$$

where

$$\tilde{s}_i := s_{2i+1}. \quad (\text{A.10})$$

By (A.5), $\tilde{s}'_\alpha(b) = \tilde{s}'_\alpha(0)(1 - b^2)^\alpha > 0$; here and in the rest of this proof, α stands for an arbitrary nonnegative integer. So, $g'(b)$ has the same sign as

$$\frac{g'(b)}{\tilde{s}'_{i+1}(b)} 2(1 - b^2)(i + 2)(2i + 3) = -1 - 2(3 + 2i)w + (15 + 16i + 4i^2)w^2 =: g_1(w) \quad (\text{A.11})$$

where $w := b^2$. Since the function g_1 is convex, with $g_1(0) = -1 < 0$ and $g_1(1) = 4(2 + 3i + i^2) > 0$, it follows that $g_1(w)$ switches exactly once in sign, from $-$ to $+$, as w increases from 0 to 1. That is, $g(b)$ switches exactly once, from decreasing to increasing, as b increases from 0 to 1. Also, by (A.1) and (A.10), $\tilde{s}_\alpha(0) = 0$ and $\tilde{s}_\alpha(1) = 1$, whence $g(0) = 0 = g(1)$. Thus, indeed $g(b) < 0$ for all $b \in (0, 1)$. \square

Lemma A.2. For $i \in \overline{0, \infty}$, let

$$C_i := \sqrt{\frac{\pi}{2}} \frac{s'_{2i+1}(0)}{\sqrt{2i+1}} = 2^{-2i} \sqrt{\pi(i + 1/2)} \binom{2i}{i}, \quad (\text{A.12})$$

the latter equality following by (A.4) and (A.2).

Then C_i decreases from $\sqrt{\frac{\pi}{2}}$ to 1 as i increases from 0 to ∞ , and for all $i \in \overline{1, \infty}$

$$C_i < \left(\frac{i+1}{i}\right)^{1/8}. \quad (\text{A.13})$$

Proof. In this proof, it is assumed that $i \in \overline{0, \infty}$. Let

$$r_i := \frac{C_i}{C_{i+1}} = \frac{2i+2}{\sqrt{(2i+2)^2 - 1}} > 1.$$

So, C_i indeed decreases in i . It is easy to check that $C_0 = \sqrt{\frac{\pi}{2}}$ and $C_i \rightarrow C_\infty := 1$ as $i \rightarrow \infty$.

It remains to verify inequality (A.13). Accordingly, assume through the end of this proof that $i \in \overline{1, \infty}$. Then

$$-1 + r_i^{-8} / \left(1 - \frac{1}{(i+1)^2}\right) = \frac{96i^4 + 384i^3 + 560i^2 + 352i + 81}{256i(i+1)^6(i+2)} > 0,$$

which shows that

$$r_i < \left(1 - \frac{1}{(i+1)^2}\right)^{-1/8},$$

whence

$$\begin{aligned} C_i &= C_\infty \prod_{\alpha=i}^{\infty} r_\alpha = \prod_{\alpha=i}^{\infty} r_\alpha < \prod_{\alpha=i}^{\infty} \left(1 - \frac{1}{(\alpha+1)^2}\right)^{-1/8} \\ &= \prod_{\alpha=i}^{\infty} \left(\frac{\alpha}{\alpha+1} / \frac{\alpha+1}{\alpha+2}\right)^{-1/8} = \left(\frac{i+1}{i}\right)^{1/8}, \end{aligned}$$

which completes the proof of Lemma A.2. \square

Lemma A.3. *Take any real $\kappa \geq 1$ and any $b \in [0, 1]$. Then*

$$\widetilde{\text{bayes}}(\kappa, b) \geq \frac{1}{2} (1 - C_{i_\kappa} \psi_b(\kappa)),$$

where $i_\kappa := \lfloor \frac{\kappa-1}{2} \rfloor$, C_i as in (A.12), and

$$\psi_b(\kappa) := \frac{\text{erf}(b\sqrt{\kappa/2})}{1 - b^2/6}.$$

Proof. For brevity, let $i := i_\kappa = \lfloor \frac{\kappa-1}{2} \rfloor$ and $k := 2i + 1$. Then $i \in \overline{0, \infty}$, $k = 2i + 1 \leq \kappa < 2i + 3 = k + 2$, and so, by (2.24),

$$\widetilde{\text{bayes}}(\kappa, b) = \frac{k+2-\kappa}{2} \text{bayes}(k, b) + \frac{\kappa-k}{2} \text{bayes}(k+2, b). \quad (\text{A.14})$$

In view of (A.2) and (A.3), $j_k = i = q_k = q$. Recalling (A.5), and using the elementary inequality $1 - t \leq e^{-t}$ for real t and the log-convexity of $\tilde{S}_q(b) := \int_0^b e^{-qu^2} du$ in real q , one has

$$\begin{aligned} S_q(b) &\leq \tilde{S}_q(b) \leq \tilde{S}_{q+1/2}(b) \frac{\tilde{S}_0(b)}{\tilde{S}_{1/2}(b)} \\ &= \frac{\sqrt{\pi}}{2} \frac{\text{erf}(b\sqrt{q+1/2})}{\sqrt{q+1/2}} \frac{b}{\int_0^b e^{-u^2/2} du} \\ &\leq \frac{\sqrt{\pi}}{2} \frac{\text{erf}(b\sqrt{q+1/2})}{\sqrt{q+1/2}} \frac{1}{1 - b^2/6} \\ &= \sqrt{\frac{\pi}{2}} \frac{1}{\sqrt{k}} \frac{1}{1 - b^2/6} \text{erf}(b\sqrt{k/2}) = \sqrt{\frac{\pi}{2}} \frac{1}{\sqrt{k}} \psi_b(k). \end{aligned} \quad (\text{A.15})$$

So, in view of (2.13), (A.5), and (A.12), $\text{bayes}(k, b) \geq \frac{1}{2} (1 - C_i \psi_b(k))$. Replacing here k by $k + 2$, one has

$$\text{bayes}(k+2, b) \geq \frac{1}{2} (1 - C_{i+1} \psi_b(k+2)) \geq \frac{1}{2} (1 - C_i \psi_b(k+2));$$

the last inequality here follows because, by Lemma A.2, C_i decreases in i . Now (A.14) yields

$$\widehat{\text{bayes}}(\kappa, b) \geq \frac{1}{2} - \frac{1}{2} C_i \left[\frac{k+2-\kappa}{2} \psi_b(k) + \frac{\kappa-k}{2} \psi_b(k+2) \right] \geq \frac{1}{2} - \frac{1}{2} C_i \psi_b(\kappa),$$

the latter inequality following by the concavity of $\psi_b(u)$ in $u \geq 0$. This completes the proof of Lemma A.3. \square

Appendix B: Details concerning $c_{m,d}^{\text{LB}}$

The second display on page 64 in [2] suggests the following lower bound on the minimax expected excess risk:

$$\text{AB}(m, d, b) := \frac{b}{4} \left(1 - \sqrt{1 - \exp \left(-\frac{(m/d+1)b^2}{1-b^2} \right)} \right).$$

Note that $1 - \sqrt{1 - e^{-x}}$ is decreasing in $x \geq 0$. Therefore, making the substitution $b = z/\sqrt{m/d+1}$ with $z \geq 0$, we see that

$$\begin{aligned} \text{AB}(m, d, b) &\leq \frac{b}{4} \left(1 - \sqrt{1 - \exp(-(m/d+1)b^2)} \right) \\ &\leq \sup_{z \geq 0} \frac{z}{4\sqrt{m/d+1}} \left(1 - \sqrt{1 - \exp(-z^2)} \right) = \frac{0.06752\dots}{\sqrt{m/d+1}} < \frac{0.06753}{\sqrt{m/d}} \end{aligned}$$

for all natural m and d . This shows that the proof of [2, Theorem 5.2] does not yield a lower bound on $c_{m,d}^{\text{LB}}$ (as in (1.10) or (2.1)) better than 0.06753.

Appendix C: Index of symbols

symbols	defined in	symbols	defined in
$\text{err}(h, D)$	(1.1)	$\text{bayes}(k, b)$	(2.13)
$\text{err}_{\min}(D)$	(1.2)	$\widehat{\text{bayes}}(\kappa, b)$	(2.24)
$\Delta(h, D)$	(1.5)	c_∞, z_*	(1.13), (2.26)
$\mathfrak{R}_m(L, D)$	(1.9)	$B_0(m, d)$	(2.19)
L_{ERM}^*	(2.10)	$B_1(\nu)$	(2.25)
$\nu = \nu_{m,d}$	(1.7)	i_ν, C_i	Thm. 2.11
$c_{m,d}^{\text{LB}}$	(2.1)	c_ν, \tilde{c}_ν	Thm. 2.11
$s_k(b)$	(2.14)	$B_2(\nu), \tilde{B}_2(\nu)$	Thm. 2.11

References

- [1] ANGLUIN, D. and VALIANT, L. G. (1979). Fast probabilistic algorithms for Hamiltonian circuits and matchings. *J. Comput. System Sci.* **18** 155–193. [MR532174](#)

- [2] ANTHONY, M. and BARTLETT, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge. [MR1741038 \(2001b:68061\)](#)
- [3] BEREND, D. and KONTOROVICH, A. (2015). A finite sample analysis of the Naive Bayes classifier. *Journal of Machine Learning Research* **16** 1519–1545.
- [4] DEVROYE, L. and LUGOSI, G. (1995). Lower bounds in pattern recognition and learning. *Pattern Recognition* **28** 1011–1018.
- [5] HAUSSLER, D. (1992). Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications. *Inf. Comput.* **100** 78–150.
- [6] HAUSSLER, D. (1995). Sphere packing numbers for subsets of the Boolean n -cube with bounded Vapnik-Chervonenkis dimension. *J. Combin. Theory Ser. A* **69** 217–232. [MR1313896 \(96f:52027\)](#)
- [7] KEARNS, M. J. and SCHAPIRE, R. E. (1994). Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.* **48** 464–497.
- [8] KEARNS, M. J., SCHAPIRE, R. E. and SELLIE, L. (1994). Toward Efficient Agnostic Learning. *Machine Learning* **17** 115–141.
- [9] LONG, P. M. (1999). The Complexity of Learning According to Two Models of a Drifting Environment. *Mach. Learn.* **37** 337–354.
- [10] NEYMAN, J. and PEARSON, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **231** 289–337.
- [11] PINELIS, I. (2016). Optimal binomial, Poisson, and normal left-tail domination for sums of nonnegative random variables. *Electron. J. Probab.* **21** 1–19.
- [12] PINELIS, I. F. (1991). Criterion for complete determinacy for concave-convexlike games. *Math. Notes* **49** 277–279.
- [13] PINELIS, I. F. and UTEV, S. A. (1989). Sharp exponential estimates for sums of independent random variables. *Theory Probab. Appl.* **34** 340–346. [MR1005745 \(91a:60053\)](#)
- [14] SHALEV-SHWARTZ, S. and BEN-DAVID, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.
- [15] SIMON, H. U. (1996). General bounds on the number of examples needed for learning probabilistic concepts. *J. Comput. System Sci.* **52** 239–254. Sixth Annual Workshop on Computational Learning Theory (COLT) (Santa Cruz, CA, 1993). [MR1393992](#)
- [16] SION, M. (1958). On general minimax theorems. *Pacific J. Math.* **8** 171–176. [MR0097026 \(20 ##3506\)](#)
- [17] TALAGRAND, M. (1994). Sharper Bounds for Gaussian and Empirical Processes. *Ann. Probab.* **22** 28–76.
- [18] V. NEUMANN, J. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* **100** 295–320.
- [19] VALIANT, L. G. (1984). A Theory of the Learnable. *Commun. ACM* **27** 1134–1142.

- [20] VAPNIK, V. N. and ČERVONENKIS, A. J. (1971). The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Veroyatnost. i Primenen.* **16** 264–279. [MR0288823](#) ([44](#) [##6018](#))